PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-110384
(43)Date of publication of application : 23.04.1999

(51)Int.Cl. G06F 17/24
G06F 17/21
G06F 17/30

(21)Application number : 10-198038
(22)Date of filing : 29.06.1998

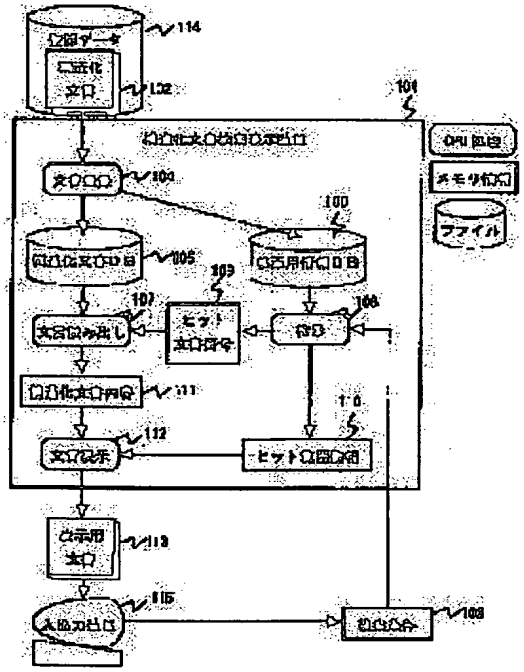
(71)Applicant : HITACHI LTD
(72)Inventor : OKAMOTO TAKUYA
MURATA HIDEKO
TAKAHASHI TORU
YAMAZAKI NORIYUKI
AOYAMA YUKI

(30)Priority

Priority number : 09190716 Priority date : 01.07.1997 Priority country : JP
09195408 22.07.1997 JP

(54) METHOD AND DEVICE FOR RETRIEVING AND DISPLAYING STRUCTURED DOCUMENT

(57)Abstract:
PROBLEM TO BE SOLVED: To retrieve a document that eliminates structure information which becomes an obstacle to retrieval and to display with highlight information added to the original structured document when displaying a retrieval result.
SOLUTION: This device performs document registration processing with a structured document 102 of a file 114 as an input, produces structured document that is undergone a structure analysis and information for document retrieval and stored them in DBs 105 and 106 respectively. Next, when an input-output device 115 inputs a retrieval condition 103, it analyzes the retrieval condition, reads information for document retrieval and performs retrieval processing 108. It outputs document number information 109 that is hit as a retrieval result and hit range information 110. Display processing first reads a corresponding structure-analyzed structured document 111 from the DB 105 based on the information 109 which is hit by document read processing 107. The processing of document display 112 embeds hit information in the document 111 based on the information 110, produces a structured document 113 for display to which the highlight information is added and shows it.



LEGAL STATUS
[Date of request for examination] 21.02.2002

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平11－110384

(43)公開日　平成11年(1999)４月23日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/24		Ｇ０６Ｆ　15/20		５５４Ｈ
17/21				５５４Ｎ
17/30				５９０Ｅ
			15/40	３４０
				３７０Ａ

審査請求　未請求　請求項の数14　ＦＤ　（全 48 頁）　最終頁に続く

(21)出願番号　特願平10－198038

(22)出願日　平成10年(1998)６月29日

(31)優先権主張番号　特願平9－190716
(32)優先日　平９(1997)７月１日
(33)優先権主張国　日本（ＪＰ）
(31)優先権主張番号　特願平9－195408
(32)優先日　平９(1997)７月22日
(33)優先権主張国　日本（ＪＰ）

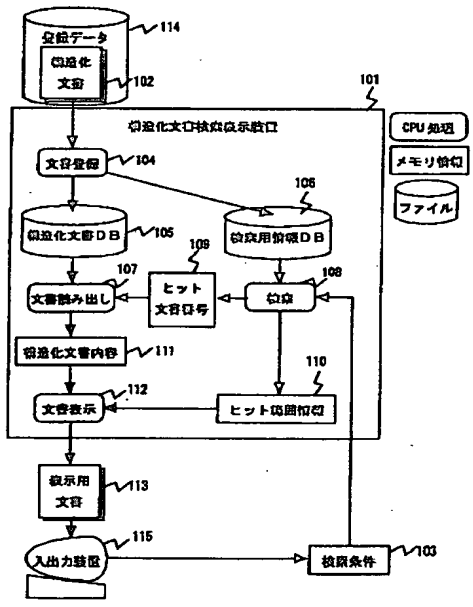
(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目６番地
(72)発明者　岡本　卓哉
神奈川県横浜市都筑区加賀原二丁目２番
株式会社日立製作所システム開発本部内
(72)発明者　村田　英子
神奈川県横浜市都筑区加賀原二丁目２番
株式会社日立製作所システム開発本部内
(72)発明者　高橋　亨
神奈川県横浜市都筑区加賀原二丁目２番
株式会社日立製作所システム開発本部内
(74)代理人　弁理士　笹岡　茂　（外１名）

最終頁に続く

(54)【発明の名称】　構造化文書検索表示方法及び装置

(57)【要約】
【課題】　検索の障害になる構造情報を除去した文書に対して検索し、検索結果の表示の際には元の構造化文書に対してハイライト情報を付加した表示をする。
【解決手段】　ファイル114の構造化文書102を入力として文書登録の処理を行い、構造解析された構造化文書と、文書検索のための情報を生成し、夫々ＤＢ105、ＤＢ106に格納する。次に入出力装置115から検索条件103が入力されると、検索条件を解析し、文書検索用情報を読み出して検索処理108を行う。検索結果としてヒットした文書番号情報109とヒット範囲情報110を出力する。表示処理は、まず、文書読み出しの処理107でヒットした文書番号の情報109に基づきＤＢ105から対応する構造解析済構造化文書111を読み出す。文書表示112の処理では、ヒット範囲情報110を基に、構造化文書111に対して、ヒット情報を埋め込み、ハイライト情報を付加した表示用の構造化文書113を生成し、これを表示する。

【図１】



<実施例１、２の処理ブロック図>

1

【特許請求の範囲】

【請求項1】　処理装置と、記憶装置と、ファイル装置と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であって、

前記処理装置は、

入力された構造化文書を解析して解析済み構造化文書を生成し、該解析済み構造化文書を前記ファイル装置に格納し、

該解析済み構造化文書から各構造内の内容文字列情報を取得して文書検索用情報を生成し、前記ファイル装置に格納し、

入力された検索条件により該ファイル装置に格納された文書検索用情報を検索し、該検索条件を満たす内容文字列情報があるか否か判定し、該検索条件を満たすとみなされる内容文字列情報を持つ文書の解析済み構造化文書を取得し、かつ該文書の検索条件を満たす範囲の情報を取得し、

該文書の検索条件を満たす範囲をハイライト表示するための表示用文書型定義（表示用ＤＴＤ）を作成し、

前記文書の検索条件を満たす範囲の情報と表示用文書型定義に基づき構造化文書中にハイライト表示するための情報を付加した表示用構造化文書を作成することを特徴とする構造化文書検索表示方法。

【請求項2】　処理装置と、記憶装置と、ファイル装置と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であって、

前記処理装置は、

入力された構造化文書を解析して解析済み構造化文書を生成し、該解析済み構造化文書を前記ファイル装置に格納し、

前記入力された構造化文書から予め与えられた検索対象外の構造情報を除去した文書検索用の正規化処理済み構造化文書を生成し、かつ該除去された構造情報を復元するための復元情報を生成し、前記ファイル装置に格納し、

入力された検索条件により該ファイル装置に格納された正規化処理済み構造化文書を検索し、該検索条件を満たす正規化処理済み構造化文書があるか否か判定し、該検索条件を満たすとみなされる文書の正規化処理済み構造化文書を取得し、かつ該文書の検索条件を満たす範囲の情報を取得し、

該文書の検索条件を満たす範囲をハイライト表示するための表示用文書型定義を作成し、

前記検索により取得された正規化処理済み構造化文書を前記復元情報により、除去された構造情報を有する構造化文書に復元し、前記文書の検索条件を満たす範囲の情報と表示用文書型定義に基づき該復元された構造化文書中にハイライト表示するための情報を付加した表示用構造化文書を作成することを特徴とする構造化文書検索表示方法。

2

【請求項3】　処理装置と、記憶装置と、ファイル装置と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であって、

前記処理装置は、

入力された構造化文書を解析して解析済み構造化文書を生成し、該解析済み構造化文書を前記ファイル装置に格納し、

該解析済み構造化文書から各構造内の内容文字列情報を取得して文書検索用情報を生成し、前記ファイル装置に格納し、

入力された検索条件により該ファイル装置に格納された文書検索用情報を検索し、該検索条件を満たす内容文字列情報があるか否か判定し、該検索条件を満たすとみなされる内容文字列情報を持つ文書の解析済み構造化文書を取得し、かつ該文書の検索条件を満たす範囲の情報を取得し、

入力された表示対象の部分構造を取得し、

該表示対象の部分構造中の前記検索条件を満たす範囲をハイライト表示するための部分構造表示用文書型定義を作成し、

該表示対象の部分構造に対して、前記文書の検索条件を満たす範囲の情報と部分構造表示用文書型定義に基づき構造化文書中にハイライト表示するための情報を付加した部分構造表示用構造化文書を作成することを特徴とする構造化文書検索表示方法。

【請求項4】　請求項1または請求項2または請求項3記載の構造化文書検索表示方法において、

検索結果のハイライト表示は、検索ターム毎に複数のハイライト表示形態のいずれかを用いてハイライト表示することを特徴とする構造化文書検索表示方法。

【請求項5】　請求項1または請求項2または請求項3記載の構造化文書検索表示方法において、

検索条件中の2つの検索タームについて、各検索タームの相対的な出現位置に関する条件を満たしている場合は、検索条件を構成する各検索タームに対するハイライト表示とその2つの検索タームを含む最小の文字列範囲に対するハイライト表示をそれぞれ異なったハイライト表示形態を用いてハイライト表示することを特徴とする構造化文書検索表示方法。

【請求項6】　請求項1または請求項2または請求項3記載の構造化文書検索表示方法において、

検索条件に複数の検索タームについて、検索条件を構成する各検索タームに対するハイライト表示と該検索タームを含む構造全体に対するハイライト表示をそれぞれ異なつた表示形態を用いてハイライト表示することを特徴とする構造化文書検索表示方法。

【請求項7】　請求項4記載の構造化文書検索表示方法において、

各検索ターム毎のハイライト表示のハイライト表示形態は、各検索タームの出現頻度の情報に基づき決定するこ

3

とを特徴とする構造化文書検索表示方法。

【請求項8】　請求項4 記載の構造化文書検索表示方法において、

各検索ターム毎のハイライト表示のハイライト表示形態は、各検索タームごとに予め与えられた重み付けの情報に基づき決定することを特徴とする構造化文書検索表示方法。

【請求項9】　処理装置と、記憶装置と、ファイル装置と、入出力装置を備える構造化文書検索表示装置であって、

前記処理装置は、

入力された構造化文書を解析して解析済み構造化文書を生成し、該解析済み構造化文書を前記ファイル装置に格納する手段と、

前記入力された構造化文書から予め与えられた検索対象外の構造情報を除去した文書検索用の正規化処理済み構造化文書を生成し、前記ファイル装置に格納する手段と、

該除去された構造情報を復元するための復元情報を生成し、前記ファイル装置に格納する手段と、

入力された検索条件により該ファイル装置に格納された正規化処理済み構造化文書を検索し、該検索条件を満たす正規化処理済み構造化文書があるか否か判定し、該検索条件を満たすとみなされる文書の正規化処理済み構造化文書を取得し、かつ該文書の検索条件を満たす範囲の情報を取得する手段と、

該文書の検索条件を満たす範囲をハイライト表示するための表示用文書型定義を作成する手段と、

前記検索により取得された正規化処理済み構造化文書を前記復元情報により、除去された構造情報を有する構造化文書に復元する手段と、前記文書の検索条件を満たす範囲の情報と表示用文書型定義に基づき該復元された構造化文書中にハイライト表示するための情報を付加した表示用構造化文書を作成する手段を有することを特徴とする構造化文書検索表示装置。

【請求項１０】　処理装置と、記憶装置と、ファイル装置と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であって、

前記処理装置は、

入力された特定の文書型定義に従う構造化文書をタグを残したままプレーンテキストとして前記ファイル装置に格納し、

入力された検索条件により該ファイル装置に格納されたプレーンテキストを検索し、該検索条件を満たす範囲があるか否か判定し、該検索条件を満たす範囲を持つ文書をプレーンテキストとして取得し、かつ該文書の検索条件を満たす範囲の情報を取得し、

前記特定の文書型定義を表示用文書型定義とし、前記入力された構造化文書に対して前記検索条件を満たす範囲に対して該表示用文書型定義に基づくハイライト表示するための情報を付加した表示用構造化文書を作成することを特徴とする構造化文書検索表示方法。

4

【請求項１１】　請求項１０記載の構造化文書検索表示方法において、

検索条件を満たす範囲が構造化文書において文書構造を示すタグの属性情報中に存在するか否かを判定し、

該検索条件を満たす範囲がタグの属性情報中に存在する場合は、構造化文書の内容文字列中に該検索条件を満たす範囲の文字列を含む文字列を追加し、該文字列において該検索条件を満たす範囲に対して前記特定の文書型定義に基づくハイライト表示するための情報を付加した表示用構造化文書を作成することを特徴とする構造化文書検索表示方法。

【請求項１２】　請求項１０記載の構造化文書検索表示方法において、

入力された検索条件により該ファイル装置にタグを残したままプレーンテキストとして格納された構造化文書を検索する際に、予め指定された特定のタグを構成する文字列を検索対象から除去し、該特定のタグを構成する文字列の前後を連結した文字列に対して検索することで得られる検索条件を満たす範囲に対して、前記特定の文書型定義に基づくハイライト表示するための情報を付加した表示用構造化文書を作成することを特徴とする構造化文書検索表示方法。

【請求項１３】　請求項１０記載の構造化文書検索表示方法において、

入力された検索条件により該ファイル装置にプレーンテキストとして格納された構造化文書を検索する際に、検索条件を満たす範囲が予め指定された文書構造の開始を示す特定のタグと文書構造の終わりを示す特定のタグに挟まれるか否かを判定し、

挟まれる場合は、文書構造の開始を示す特定のタグより前もしくは文書構造の終わりを示すタグより後ろの内容文字列中に、該検索条件を満たす範囲の文字列を含む文字列を追加し、該文字列において該検索条件を満たす範囲に対して前記特定の文書型定義に基づくハイライト表示するための情報を付加した表示用構造化文書を作成することを特徴とする構造化文書検索表示方法。

【請求項１４】　請求項1または請求項2または請求項3または請求項１０記載の構造化文書検索表示方法において、

前記表示用文書型定義に基づくハイライト表示するための情報を付加した表示用構造化文書を作成する際に、ハイライト表示するための情報は、検索条件中に指定された方法を用いて付加することを特徴とする構造化文書検索表示方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、ＳＧＭＬ、ＨＴＭＬなどによって作成された構造化文書に対する検索表示

5

技術に係り、特に構造化文書に対して検索を行い、検索結果に対してハイライトして表示する構造化文書検索表示方法および装置に関する。

【０００２】

【従来の技術】ワードプロセッサなどの普及により、作成される文書情報の電子化が進んでいる。これらの電子化文書は、作成される機器、ソフトウェアによって個々のフォーマットを持っており、別の機器あるいはソフトウェアでは、利用できない、あるいは、何らかの変換手段を用意することが必要となっていた。このような文書交換のための共通フォーマットとして、各種の構造化文書が提案されている。これらの構造化文書は、文書の基本構造である、章、節、項などの階層構造を定義できるだけでなく、レイアウト情報を含む事も可能となっている。

【０００３】構造化文書の記述言語として、標準化が進められているのが、ＳＧＭＬ（Standard Generalized Markup Language）＝「標準一般化マークアップ言語」である。ＳＧＭＬは、構造化文書の構造情報をタグと呼ばれる特定の文字列をテキスト中に埋め込むことで、文書の構造を表現する方法を用いている。ＳＧＭＬでは、タグの名称、内容、さらに、タグによって示される文書構造をＤＴＤ（DocumentType Definition）＝「文書型定義」によって規定することができる。上記のＳＧＭＬ、ＤＴＤについては、「実践ＳＧＭＬ」（ＳＧＭＬ懇談会実用化WG監訳　1992年4月20日 財団法人 日本規格協会発行）に詳細に説明されている。これらの構造化文書を検索システムのＤＢに登録して、構造名を指定して検索しようとする場合を想定する。登録しようとする各文書のＤＴＤが異なる場合、処理方法としては、文書ごとに文書構造を解析して、指定された構造名がどの部分に相当するかを解析した上で、検索対象の文字列を取得して検索する方法が考えられる。しかし、この方法は、多くの処理時間を必要とする。また、構造名ごとに各文書の対応する箇所をテーブルで持つなどの方法を用いる場合、各文書に出現する構造名を全て一括して管理し、構造名ごとに各文書の対応する部分を登録する必要があり、膨大な管理テーブルが必要となる。さらに、異なるＤＴＤが混在する文書を登録しても、検索対象の構造をすべての文書が持っているとは限らず、また、例えば、「要約」、「要旨」のように、同じ内容であっても異なる構造名を付けた場合、これらの異なる構造名を全て指定して、検索を行なわなければならず、現実的な構造化文書の検索とは考えられない。

【０００４】したがって、同じ文書型定義で生成された文書だけを登録するように運用することが構造化文書の検索では必要となる。あらかじめ指定された構造名について、各文書の対応する部分を管理する。検索の際には、検索対象の構造名および検索条件を指定すると、各文書の指定された構造に対応する部分に検索条件に当てはまる文字列が含まれると、検索条件にヒットしたと判断される。

6

【０００５】構造化文書の検索結果として、文書の内容を表示するための機能の従来技術について以下に述べる。まず、第1の従来技術として、特開平8-339369「文書表示装置および文書表示方法」が挙げられる。本従来技術は、ＳＧＭＬ文書の構造解析および構造表示用のレイアウトへの変換、さらに指定構造の内容の表示を行う方法について述べられており、本技術を用いることで構造化文書を構造単位で表示することが可能である。さらに、本従来技術においては、指定構造のハイライト表示（強調した表示のことであり、色、字体、字の大きさ等を変えたり、アンダーラインを付したりする）の手段を提供している。しかし、ここで示されているハイライト表示手段とは、構造毎に表示方法をコントロールする手段であり、構造単位に、表示の有無、ハイライト表示などの指定を行う。したがって、本従来技術において、構造化文書の検索結果の表示を実現する際に必要となる、ヒットした検索タームに対するハイライト表示を実現する方法が示されているわけではない。

【０００６】また、第2の従来技術としては、特開平8-212230「文書検索方法および文書検索装置」で構造化文書以外の文書の検索結果に対するハイライト表示方法が示されている。しかし、本従来技術は、表示するためのテキストに対するヒット範囲の取得およびハイライト情報の付加を実現するのみであり、構造化文書の検索結果として得られた文書に対してハイライト情報を付加する機能を持つわけではない。

【０００７】上記2つの従来技術を組み合わせただけでは、構造化文書に対する検索結果として出力する文書に対して、ヒットしたタームに対するハイライト情報の付加を実現する事はできない。つまり、構造化文書において、ハイライト表示を実現するためには、表示対象の文書の作成時のＤＴＤにハイライト用の構造情報を追加したＤＴＤを作成する手段が必要となる。

【０００８】構造化文書にハイライト情報を付加した際の文書型定義の変更方法については、第3の従来技術である、特願平8-159202「構造化文書の版管理方法および装置」に、元のＤＴＤに対して新たな構造を追加したＤＴＤを生成する方法が示されている。本従来技術を用いる事により、ハイライト情報を付加した文書型定義を作成することができる。

【０００９】第1、第2の従来技術により、構造化文書を構造が分かるように表示すること、さらに構造化されていない文書においては、ヒット範囲のハイライト表示をする事が可能であることがわかる。さらに、第3の従来技術を用いることにより、構造ごとに取得したハイライト情報を付加した文書型定義が指定できる。これらの技術を組み合わせることで、構造化文書の特定の構造の検索結果に対してハイライト情報を付加した構造化文書

を出力し、ハイライト表示を実現する事ができる。

【００１０】また、最新の情報を入手する方法として、近年インターネットが爆発的に広まっている。インターネット上に存在する数多くの情報から自分が必要とする情報をいち早く知る手段として、Web上の情報の検索機能も充実してきた。ＨＴＭＬ（Ｈyper Ｔext Markup Ｌanguage）は、WWW（World Wide Web）上において、文書内容を記述し、他の資源へのリンク情報、文書のフォーマットを表現するための言語である。ＨＴＭＬは、特定のＤＴＤにしたがって記述されたＳＧＭＬとみなすことができる。このＨＴＭＬ文書を作成、加工する手段として、ＨＴＭＬエディタがある。また、作成されたＨＴＭＬ文書を解析し、表示するＨＴＭＬブラウザが存在する。ＨＴＭＬブラウザには、検索する文字列（以下、「検索ターム」という。）を入力し、表示中のＨＴＭＬ文書に対して検索を行い、ヒットした箇所を反転表示などの強調表示を行う機能を持つものがある。ＳＧＭＬについても、レイアウト表示し、加工する機能を持つＳＧＭＬブラウザが存在する。ＳＧＭＬブラウザには、ブラウザ上に表示中のＳＧＭＬ文書に対して、全文検索し、検索条件に適合する箇所をハイライト表示する。これらのブラウザでは、文書表示の際に文書の解析を行ない、表示用のデータを作成している。検索はこのブラウザ上の表示用のデータに対して検索を行ない、画面上でヒット位置をハイライト表示している。

【００１１】

【発明が解決しようとする課題】上記の従来技術の組み合わせにより、与えられた構造化文書に対して、構造毎に検索した結果を、個々にハイライト表示することが可能である。しかし、構造情報には、章、節、項のように文書構造そのものを表わしているものだけでなく、アンダーラインの付加などレイアウト用の情報も含まれる場合がある。これらの構造情報は、必ずしも文の切れ目で挿入されるとは限らない。文書検索する際には、このような構造情報を除去しなければ、文書中に含まれている語であるにもかかわらず、検索できないという問題がある。このように、検索時に不要となる構造情報を除去する処理を、以下の説明では「正規化処理」と呼ぶ。正規化処理を行なった構造化文書を検索対象とし、元の構造化文書に対してハイライト情報を付加した表示を実現するためには、正規化処理を行なった構造化文書に対して、上記の従来技術を用いた方法を利用するだけでは実現できない。つまり、この方法では、検索時には、元の文書の構造情報の一部しか残っていないため、この構造情報に対してハイライト情報を付加するだけでは、元の構造化文書に対してヒットした検索タームのハイライト表示を実現することにならないのである。

【００１２】一方、ＨＴＭＬ文書は、ブラウザ依存の独自の拡張により複数のＤＴＤに基づいて作成されたＨＴＭＬ文書が存在し、またどのＤＴＤに基づいて記述されているかがわからない。さらに、ＳＧＭＬの文法に基づいて正しく記述されていない文書も数多く存在するため、ＳＧＭＬと同様の方法で構造解析することは困難である。また、（１）プレーンなテキスト文書に対しては、検索処理を行い、検索ヒット位置の前後にハイライト用のタグを挿入したＨＴＭＬ文書を生成することにより、ＨＴＭＬブラウザ上で、検索ヒットした文字列を強調表示することが可能である。しかし、タグ内の文字列が検索タームと一致した場合、この検索ヒット位置の前後に対して、ハイライト用のタグを挿入すると、元々のＨＴＭＬのタグの内容が変更されるため、正しく表示されなくなるといった問題が起こる。さらに、（２）ＨＴＭＬブラウザ上で連続して表示されている文字列の途中に、レイアウトを表現するタグが挿入されている場合があり、ＨＴＭＬ文書に対して検索する場合は、タグを除いて検索しなければ正しく検索することができない。例えば、ＨＴＭＬ文書中に「今月の＜ＦＯＮＴ　ＳＩＺＥ＝＋１＞特集＜／ＦＯＮＴ＞記事」と書かれており、検索タームを「特集記事」とした場合、ＨＴＭＬ文書中では、「特集」と「記事」の間に文字を拡大して表示するための「＜／ＦＯＮＴ＞」のタグが記述されているため、タグを飛ばして検索しなければ正しく検索することができない。

【００１３】本発明の目的は、正規化処理された文書に対する検索結果から、元の文書に対するハイライト情報の付加を実現するために、検索用の文書から、元の文書のハイライト範囲情報への変換を実現することにある。本発明の他の目的は、正規化後のヒットタームが、元の文書において複数の構造にまたがっている場合、各構造ごとに、ヒットした範囲に対してハイライト情報を付加し、ハイライト表示することにある。本発明のさらに他の目的は、ヒットしたタームが含まれる構造全体のハイライト表示、あるいは、出現位置の距離条件を満たした２つの検索タームを含む領域全体をハイライト表示するなどの処理をするため、階層的なハイライト情報を付加し、異なるハイライト表示形態によりハイライト表示することにある。本発明のさらに他の目的は、構造化文書の部分構造だけを抽出して表示する場合に、このような部分構造の内容についても、ハイライト情報を付加し、ハイライト表示することにある。本発明のさらに他の目的は、文書構造を示すＨＴＭＬタグが存在する文書から文字列を検索する場合、設定した検索タームと一致した文字列がＨＴＭＬタグ内に存在する場合や、検索タームがＨＴＭＬタグをまたがって記述されている場合でも検索を可能にすることにある。本発明のさらに他の目的は、検索条件にヒットした文字列をハイライト表示可能にすることにある。

【００１４】

【課題を解決する為の手段】上記の課題を解決するため、本発明は、処理装置と、記憶装置と、ファイル装置

と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であり、前記処理装置は、入力された構造化文書を解析して解析済み構造化文書を生成し、該解析済み構造化文書を前記ファイル装置に格納し、該解析済み構造化文書から各構造内の内容文字列情報を取得して文書検索用情報を生成し、前記ファイル装置に格納し、入力された検索条件により該ファイル装置に格納された文書検索用情報を検索し、該検索条件を満たす内容文字列情報があるか否か判定し、該検索条件を満たすとみなされる内容文字列情報を持つ文書の解析済み構造化文書を取得し、かつ該文書の検索条件を満たす範囲の情報を取得し、該文書の検索条件を満たす範囲をハイライト 表示するための表示用文書型定義(表示用ＤＴＤ) を作成し、前記文書の検索条件を満たす範囲の情報と表示用文書型定義に基づき構造化文書中にハイライト 表示するための情報を付加した表示用構造化文書を作成するようにしている。

【００１５】また、処理装置と、記憶装置と、ファイル装置と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であり、前記処理装置は、入力された構造化文書を解析して解析済み構造化文書を生成し、該解析済み構造化文書を前記ファイル装置に格納し、前記入力された構造化文書から予め与えられた検索対象外の構造情報を除去した文書検索用の正規化処理済み構造化文書を生成し、かつ該除去された構造情報を復元するための復元情報を生成し、前記ファイル装置に格納し、入力された検索条件により該ファイル装置に格納された正規化処理済み構造化文書を検索し、該検索条件を満たす正規化処理済み構造化文書があるか否か判定し、該検索条件を満たすとみなされる文書の正規化処理済み構造化文書を取得し、かつ該文書の検索条件を満たす範囲の情報を取得し、該文書の検索条件を満たす範囲をハイライト 表示するための表示用文書型定義を作成し、前記検索により取得された正規化処理済み構造化文書を前記復元情報により、除去された構造情報を有する構造化文書に復元し、前記文書の検索条件を満たす範囲の情報と表示用文書型定義に基づき該復元された構造化文書中にハイライト 表示するための情報を付加した表示用構造化文書を作成するようにしている。

【００１６】また、処理装置と、記憶装置と、ファイル装置と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であり、前記処理装置は、入力された構造化文書を解析して解析済み構造化文書を生成し、該解析済み構造化文書を前記ファイル装置に格納し、該解析済み構造化文書から各構造内の内容文字列情報を取得して文書検索用情報を生成し、前記ファイル装置に格納し、入力された検索条件により該ファイル装置に格納された文書検索用情報を検索し、該検索条件を満たす内容文字列情報があるか否か判定し、該検索条件を満たすとみなされる内容文字列情報を持つ文書の解析済み構造化文書を取得し、かつ該文書の検索条件を満たす範囲の情報を取得し、入力された表示対象の部分構造を取得し、該表示対象の部分構造中の前記検索条件を満たす範囲をハイライト 表示するための部分構造表示用文書型定義を作成し、該表示対象の部分構造に対して、前記文書の検索条件を満たす範囲の情報と部分構造表示用文書型定義に基づき構造化文書中にハイライト 表示するための情報を付加した部分構造表示用構造化文書を作成するようにしている。

【００１７】また、処理装置と、記憶装置と、ファイル装置と、入出力装置を備える構造化文書検索表示装置であり、前記処理装置は、入力された構造化文書を解析して解析済み構造化文書を生成し、該解析済み構造化文書を前記ファイル装置に格納する手段と、前記入力された構造化文書から予め与えられた検索対象外の構造情報を除去した文書検索用の正規化処理済み構造化文書を生成し、前記ファイル装置に格納する手段と、該除去された構造情報を復元するための復元情報を生成し、前記ファイル装置に格納する手段と、入力された検索条件により該ファイル装置に格納された正規化処理済み構造化文書を検索し、該検索条件を満たす正規化処理済み構造化文書があるか否か判定し、該検索条件を満たすとみなされる正規化処理済み構造化文書の情報を取得し、かつ該文書の検索条件を満たす範囲の情報を取得する手段と、該文書の検索条件を満たす範囲をハイライト 表示するための表示用文書型定義を作成する手段と、前記検索により取得された正規化処理済み構造化文書を前記復元情報により、除去された構造情報を有する構造化文書に復元する手段と、前記文書の検索条件を満たす範囲の情報と表示用文書型定義に基づき該復元された構造化文書中にハイライト 表示するための情報を付加した表示用構造化文書を作成する手段を有するようにしている。

【００１８】また、処理装置と、記憶装置と、ファイル装置と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であり、前記処理装置は、入力された特定の文書型定義に従う構造化文書をタグを残したままプレーンテキストとして前記ファイル装置に格納し、入力された検索条件により該ファイル装置に格納されたプレーンテキストを検索し、該検索条件を満たす範囲があるか否か判定し、該検索条件を満たす範囲を持つ文書をプレーンテキストとして取得し、かつ該文書の検索条件を満たす範囲の情報を取得し、前記特定の文書型定義を表示用文書型定義とし、前記入力された構造化文書に対して前記検索条件を満たす範囲に対して該表示用文書型定義に基づくハイライト 表示するための情報を付加した表示用構造化文書を作成するようにしている。

【００１９】また、処理装置と、記憶装置と、ファイル装置と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であり、前記処理装置は、入力された特定の文書型定義に従う構造化文書をタグを残し

11

たままプレーンテキストとして前記ファイル装置に格納し、入力された検索条件により該ファイル装置に格納されたプレーンテキストを検索し、該検索条件を満たす範囲があるか否か判定し、該検索条件を満たす範囲を持つ文書をプレーンテキストとして取得し、かつ該文書の検索条件を満たす範囲の情報を取得し、検索条件を満たす範囲が構造化文書において文書構造を示すタグの属性情報中に存在するか否かを判定し、該検索条件を満たす範囲がタグの属性情報中に存在する場合は、構造化文書の内容文字列中に該検索条件を満たす範囲の文字列を含む文字列を追加し、該文字列において該検索条件を満たす範囲に対して前記特定の文書型定義に基づくハイライト表示するための情報を付加した表示用構造化文書を作成するようにしている。

【００２０】また、処理装置と、記憶装置と、ファイル装置と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であり、前記処理装置は、入力された特定の文書型定義に従う構造化文書をタグを残したままプレーンテキストとして前記ファイル装置に格納し、予め指定された特定のタグを構成する文字列を検索対象から除去し、該特定のタグを構成する文字列の前後を連結した文字列に対して検索することで得られる検索条件を満たす範囲に対して、前記特定の文書型定義に基づくハイライト表示するための情報を付加した表示用構造化文書を作成するようにしている。

【００２１】また、処理装置と、記憶装置と、ファイル装置と、入出力装置を備える情報処理システムにおける構造化文書検索表示方法であり、前記処理装置は、入力された特定の文書型定義に従う構造化文書をタグを残したままプレーンテキストとして前記ファイル装置に格納し、入力された検索条件により該ファイル装置にプレーンテキストとして格納された構造化文書を検索する際に、検索条件を満たす範囲が予め指定された文書構造の開始を示す特定のタグと文書構造の終わりを示す特定のタグに挟まれるか否かを判定し、挟まれる場合は、文書構造の開始を示す特定のタグより前もしくは文書構造の終わりを示すタグより後ろの内容文字列中に、該検索条件を満たす範囲の文字列を含む文字列を追加し、該文字列において該検索条件を満たす範囲に対して前記特定の文書型定義に基づくハイライト表示するための情報を付加した表示用構造化文書を作成するようにしている。

【００２２】

【発明の実施の形態】第1の実施例の概略の処理ブロック図を図1に示す。１０１は、構造化文書検索表示装置である。登録データファイル（１１４）に格納された、構造化文書（１０２）を入力として文書登録の処理を行う事で、構造解析された構造化文書（図3により後述する）と、文書検索のための文書検索用情報（図5により後述）が生成される。構造解析された構造化文書は、構造化文書データベース（以下、データベースをＤＢと記

12

述する。）（１０５）に格納し、検索用情報は、検索用情報ＤＢ（１０６）に格納される。次に入出力装置（１１５）から、検索条件（１０３）が入力されると、検索条件を解析し、文書検索用情報を読み出して、検索処理（１０８）を行う。検索結果としては、ヒットした文書番号の情報（１０９）とヒット範囲の情報（１１０）を出力する。表示処理は、まず、文書読み出しの処理（１０７）で、ヒットした文書番号の情報（１０９）に基づいて、構造化文書ＤＢ（１０５）から、指定された構造解析済構造化文書（１１１）を読み出す。文書表示（１１２）の処理では、ヒット範囲情報（１１０）を基に、構造解析済構造化文書（１１１）に対して、ヒット情報を埋め込んだ表示用の構造化文書（１１３）を生成する。生成された表示用の構造化文書は、入出力装置（１１５）に表示される。

【００２３】図2に構造化文書検索表示の処理フローを示す。まず、構造化文書の登録処理を行なう（２０１）。登録処理の内容については、図4のフローチャートを用いて後述する。次に、指定された検索条件を用いて構造化文書を検索する（２０２）。検索処理の詳細は、図6のフローチャートを用いて後述する。検索結果としては、ヒット文書数とヒット文書を識別する番号と各文書毎の検索タームのヒット範囲がある。ヒット範囲の情報は、ヒットした検索タームが含まれる構造を識別するための構造ＩＤ（構造識別子）と構造内でのヒット開始位置、テキスト長の情報を出力する。構造化文書検索の処理で、ヒット文書数が1以上であれば（２０３）、順次、ヒットした文書の内容を読み出し（２０４）、読み出した文書のヒット範囲情報を取得し（２０５）、ハイライト表示を実現する（２０６）。表示処理の詳細については、図9を用いて後述する。さらにヒットした文書があれば、２０４から２０６の処理を繰り返す。表示処理を終えると、次の検索処理の有無を確認し（２０８）、検索条件がなければ、処理を終え、検索条件があれば、２０２の処理に戻って構造化文書の検索表示処理を繰り返す。

【００２４】図3は、構造化文書登録処理の概要を示した図である。まず、ＳＧＭＬ文書（３０１）の構造を解析し、木構造（３０２）を生成する。生成した木構造の各項目の内容をテーブル形式のデータ（３０３）として出力し、これを解析済み構造化文書として登録する。ここで、CDATAとは文字列データのことである。

【００２５】図4は構造化文書登録処理のフローチャートである。まず、構造化文書を解析する（４０１）。解析された構造化文書を解析済構造化文書として登録する（４０２）。構造化文書の解析には、ＤＴＤを利用してＳＧＭＬ文書を解析するＳＧＭＬパーサを用いることで実現できる。次に、解析された構造化文書に対して、検索に不要な構造を除去するための正規化処理を行なう（４０３）。正規化処理の手順については、図１２を用

いて後述する。そして、正規化処理した構造化文書を、文書データベースに登録する（４０４）。さらに、データベースに登録された解析済み構造化文書から、構造化文書の検索に必要な検索用情報として、構造情報、構造内のテキストの情報を取り出す（４０５）。ここで得られた検索用情報を検索用情報ＤＢ（１０６）に登録する（４０６）。ここで、登録される検索用情報は、ＳＧＭＬ文書中の構造情報（タグ）を除去し、各構造ごとに構造情報とその内容を表すテキスト列を格納したものである。図5に上記検索用情報と正規化した構造化文書からなる検索用のテキストの格納例を示す。上記処理を登録文書に対して繰り返し実行し、登録文書が無くなったとき処理を終了する（４０７）。登録内容は登録文書の全文検索に用いる。図5は、検索用のテキストとして、出力される内容の例である。このように文書構造の構造ＩＤとテキスト列を対応付けるテーブルと文字列情報からなる情報を検索用のテキストとして登録する。検索の際には、構造ＩＤを元に必要な文字列を抽出して検索を行なう。

【００２６】図6は、図2の構造化文書検索表示処理の２０２ステップの構造化文書検索の処理フローである。検索条件は、「検索対象の構造指定：検索条件式」のように与えられる。検索対象の構造は、例えば、「<文書.タイトル>」のように、'<'と'>'で囲まれ、上位構造（例の場合、「文書」）と下位構造（例の場合、「タイトル」）は'.'で区切られ、階層構造中のどの構造に対して、検索を行なうかが指定される。検索条件式は、例えば、and("検索","文書")では、"検索"と"文書"が両方出現する条件を示しており、C<=10("検索","文書")では、"検索"と"文書"が10文字以下の文字を挟んで出現する条件を示している。

【００２７】構造化文書検索は、まず、ヒット文書数のカウンタをクリアし（６０１）、次に、検索条件中の検索対象の構造指定の部分の解析を行なう（６０２）。ステップ６０２では、<文書.タイトル>のように構造を指定する文字列から、解析済み構造化文書の対応する構造を一意に特定できる構造ＩＤ（構造識別子）を取得する。構造ＩＤ取得の処理内容は、図7のフローチャートを用いて後述する。次に、検索対象として登録された文書（検索用のテキスト）を読み出し、ステップ６０２で取得した指定構造ＩＤに対応するテキスト部分を取得する（６０３）。検索条件から、検索ターム、さらに複数の検索タームの出現の論理積、距離条件などの論理条件からなる検索条件式を解析し（６０４）、得られた検索タームによりステップ６０３で取得されたテキスト部分の全文検索を行ない、検索条件式の論理条件を満たすか否かの判定、すなわち、検索条件にヒットしたか否か判定する（６０５）。検索条件にヒットすると（６０６）、検索結果として文書の番号、検索タームが含まれる構造のＩＤと、構造中の検索タームがヒットした範囲

の情報を出力する（６０７）。さらに、ヒットした文書の数をカウントし（６０８）、本処理を全文書について行なった後（６０９）、ヒット文書数を出力する（６１０）。

【００２８】図7は、図6の検索条件の解析における、構造指定内容の解析処理のフローチャートである。まず、文書の最上位構造を取得する（７０１）。次に最上位構造から順に下位構造を取得する。取得した構造が指定構造の下位構造であれば（７０３）、その構造を検索対象の構造として構造ＩＤを出力する（７０４）。下位構造があれば（７０５）、さらにその下位構造に対して、同様に指定された構造の下位構造か否かを判定し、下位構造であれば構造ＩＤを出力する処理（７０６）を下位構造がなくなるまで繰り返し（７０７）、全ての構造について処理が終われば、検索対象の構造ＩＤの一覧が得られる。図8に検索対象となる構造ＩＤ一覧の出力形式を示す。検索対象となる構造ＩＤの数（８０１）と、検索対象として得られた数のＩＤ（８０２）が出力される。

【００２９】図9は、表示処理の内容を示すフローチャートである。本フローチャートを用いて、表示処理の内容を以下に述べる。まず、検索対象の構造化文書は、検索に不要な構造を除去する正規化処理を行なった後の文書であるため、検索によりヒットした構造およびヒット範囲情報は、必ずしも登録した正規化していない構造化文書における構造および範囲と一致するとは限らない（図3の木構造３０２と図１２の木構造１２０３を参照）。表示に用いる文書は、登録した正規化していない構造化文書に対して、ヒットした範囲にハイライト情報を付加した文書となる。したがって、まず、登録文書のＤＴＤから、表示に用いる文書用の表示用ＤＴＤの作成処理を行なう（９０１）。表示用ＤＴＤ作成処理の内容については、図１１を用いて後述する。さらに、正規化後の構造化文書に対して得られたヒット範囲については、正規化前の登録した構造化文書における構造およびハイライト範囲情報に変換する（９０２）。正規化後の文書のヒット範囲情報の正規化前の文書のハイライト範囲情報への変換処理の内容については、図１５を用いて後述する。

【００３０】次に表示に用いる解析済み文書の最上位構造の情報を読み出し、９０３から９１１の処理を順に繰り返すことで、表示用の文書の出力処理を行なう。まず、構造情報を読み出し（９０３）、最初に構造の開始タグを出力する（９０４）。さらに本構造に下位構造が存在するなら（９０５）、下位構造に対して、表示処理（９０３から９１１の処理）を再帰的に行なう（９０６）。下位構造がなくなれば、構造の終わりを示すタグを出力する処理（９１１）に移る。

【００３１】ここで、下位構造とは、文字列を含む。したがって、

15

<文書>
<タイトル>
構造化文書
</タイトル>
<本文>
<強調>構造化文書</強調>の検索は、・・・
</本文>
</文書>

などの構造化文書については、<タイトル>の下位構造として、文字列（ＳＧＭＬでは、CDATAと表現される）という構造が存在することになる。CDATAは、下位構造を持たず、文字列情報として、上記の例の場合、「構造化文書」という内容を持つのである。<本文>についても同様に、<強調>という構造と、「の検索は、・・・」という内容を持つ文字列が下位構造として存在することになる。

【００３２】９０５のステップで下位構造が存在しないと判定された場合は、文字列の構造であるため、本構造の内容に対して、ヒット範囲情報と比較し（９０８）、ヒット範囲が含まれる構造であれば、ハイライト処理を行なう（９０９）。ハイライト処理については、図１６を用いて後述する。ヒット範囲が含まれない文字列であれば、内容をそのままテキストとして出力する（９１０）。出力内容が文字列の場合は、９０４、９１１のステップで、開始タグ、終了タグは出力しない。上記の処理で構造ごとのハイライト表示を実現する。さらに処理すべき構造があれば、９０３からの処理を繰り返す（９１２）。

【００３３】図１０は、登録用ＤＴＤ（１００１）と、登録するＳＧＭＬ文書（文書インスタンス）の例（１００２）、ハイライト表示に用いる表示用ＤＴＤ（１００３）と、表示用に変換したＳＧＭＬ文書（文書インスタンス）の例（１００４）である。なお、ＤＴＤ（Document Type Definition）とは、従来の技術の項で述べたように、タグの名称、内容、さらに、タグによって示される文書構造を規定する文書型定義である。ＤＴＤにおいて、構造を表現する場合は、"<!ELEMENT タグ名"に続いて、"-"または"O"が2つ並べられる。最初の"-"または、"O"は、構造開始タグの省略の可否を示しており、"-"の場合は、省略できない。"O"の場合は省略可能である。２つめの"-"または"O"は、終了タグの省略の可否を示している。次に、内容モデルとして、下位構造に出現しうる構造が記述される。図１０のＤＴＤ１００１の（タイトル，本文）の場合、タイトルは下位構造1、本文は下位構造2である。"(下位構造1 ,下位構造2 ?)"のように記述される場合は、下位構造1の後に下位構造２がそれぞれ1回だけ出現することを示し、"?"は、下位構造2は、出現しなくても良いことを示している。"(下位構造1 |下位構造2)*"の場合は、下位構造1、２が順序不同で複数回（０回を含む）出現することを示す。ここ

16

で、内容モデルに"CDATA"と記述されている場合は、その構造中には、１つだけの文字列が存在することを示している。#PCDATAも文字列を表わしているが、繰り返し出現が可能である。文字列と、構造が混在する場合は、#PCDATAを用いる必要がある。

【００３４】内容モデルに、"CDATA"の代わりに"RCDATA"が指定される場合がある。CDATAとRCDATAの違いは、CDATAが、構造内にエンティティ参照（"<xxxx;"のように記述される。外字への置き換えなどに利用される。）が出現した場合に、エンティティ（外字など）への変換を行なわないで、出現した文字列のまま、文字列として扱うのである。"RCDATA"が指定された場合は、エンティティへの変換を行なった文字列を、文字列として扱う。

【００３５】ハイライト表示するためには、文字列に対してハイライト情報を付加できるように、文書構造を変更する必要がある。１００３にアンダーラインで示した変更点のように、各構造の文字列部分に対しては、全てハイライト表示用の構造情報を追加し、さらにハイライト表示用の構造情報（<!ELEMENT ハイライト　　- -(#PCDATA)>を付加する必要がある。元のＤＴＤで内容モデルの"CDATA"となっている部分が、"(#PCDATA|ハイライト)*"に変更されているのは、CDATAがその構造中には、文字列が1つしか存在しないことを示しており、繰り返しの要素としては出現し得ないためである。ハイライト用のタグが付加されるため、元の構造がCDATAであっても、#PCDATAに変更した上で、ハイライトが繰り返し出現することが可能なように、"(#PCDATA|ハイライト)*"とするのである。

【００３６】図１１は、登録用のＤＴＤからハイライト表示用のＤＴＤを作成するための処理内容を表すフローチャートである。まず、登録用ＤＴＤを読み出し（１１０１）、ＤＴＤの内容を解析して、ELEMENT項目を取得する（１１０２）。ELEMENT項目の内容モデル中に、CDATA、RCDATA、#PCDATAなどが指定されている場合は、全て、ハイライト用の構造を付加できるように、内容モデルを変更する（１１０３－１１０６）。内容モデルの変更は、まず、"CDATA"、"RCDATA",を"#PCDATA"に変更した上で、"#PCDATA"を"(#PCDATA|ハイライト)*"のように、ハイライトタグで囲まれた文字列と、囲まれていない文字列が繰り返し出現するように定義する。元の内容モデルが、"(#PCDATA|アンダーライン)*"のように複数の構造が、繰り返し出現するように記述されている場合は、"(#PCDATA|アンダーライン|ハイライト)*"のように、ハイライト構造が出現することを記述するだけで良い。すべてのELEMENT宣言について変更処理が終わると（１１０７）、ハイライト用の構造の定義として、"<!ELEMENT ハイライト　- - CDATA>"を追加する（１１０８）。以上の処理で、図１０の１００３に示したハイライト表示用のＤＴＤが生成される。

【００３７】図12は、構造化文書の正規化処理の内容

を示した図である。図10の1001に示した構造化文書を木構造に表わすと1201のようになる。不要な構造として"アンダーライン"が指定されている場合、正規化処理の最初の処理として、1202に示すように、アンダーラインという構造を削除し、アンダーラインの下位構造に含まれる文字列は、直接上位構造である"本文"の要素とする。さらに、"本文"の下位構造として、文字列(CDATA)が2つ並んでいるため、1203のように、文字列を連結して、1つの文字列データとする。

【0038】図13は、正規化処理前の構造化文書(1301)、正規化処理後の構造化文書(1302)の内容を解析し、テーブル形式に変換して出力した内容である。1303は、構造情報を格納したテーブルであり、0から6までの構造ID が付けられた構造は、正規化前の構造の情報である。0が最上位構造であり、下位構造の情報をたどっていくことで、文書構造が分かる。7から9までの構造ID(構造識別子)が付けられた構造は、正規化後に変更、追加された構造である。7が最上位構造であり、下位構造を辿ると正規化後の文書構造が分かる。ここで、変更のない構造である"タイトル"以下の構造である構造ID 1,2の構造情報はそのまま残される。さらに、正規化処理で追加された構造ID 7から9の構造については、1304の正規化対応テーブルにより、正規化前の構造との対応関係が格納される。

【0039】図14は、正規化後の構造化文書に対して、検索した際のヒット範囲の情報を正規化前の構造化文書における範囲情報へ変換した結果を示している。1401の正規化後の構造情報に基づいて得られたヒット範囲の情報を、図13の1304の正規化対応テーブルの情報を利用して、正規化前の構造化文書における範囲情報(1402)に変換している。本図の例では、正規化後の構造ID 9のヒット範囲が、正規化前の文書では、構造ID 5と6に分かれているため、2つの構造中のハイライト対象の範囲情報に変更している。

【0040】図15に、図9の902ステップの正規化処理後の構造化文書に対するヒット範囲情報を正規化処理前の構造化文書に対するヒット範囲情報に変換する処理内容のフローチャートを示す。まず、正規化後のヒット範囲情報を順次読み出し(1501)、ヒット範囲情報の構造ID が、正規化後に追加されたものか、正規化前から存在するものであるかを判定する(1502)。正規化前から存在する構造ID であれば変更はないため、そのまま、正規化前のヒット範囲情報として出力する(1503)。正規化後に作成された構造ID であれば、図14の正規化対応テーブルの正規化後構造ID を辿り、文字範囲の情報から、対応する正規化前の構造ID と、ヒット範囲を得る(1504)。正規化処理前の構造におけるヒット範囲を得たら、これを正規化前のヒット範囲として出力する(1505)。全てのヒット範囲情報について処理を終える(1506)と、表示用のハイライト範囲情報が得られる。

【0041】図16は、図9の909ステップのハイライト処理のフローチャートである。まず、文書の先頭から、ハイライト開始までの文字列を出力する(1601)。次に、ハイライト表示に用いる構造の開始タグを出力する(1602)。さらに、ハイライト範囲の文字列を出力し(1603)、ハイライト表示に用いる構造の終了タグを出力する(1604)。すべてのハイライト処理を終えると(1605)、残ったテキストを出力し、ハイライト処理を終わる(1606)。

【0042】次に第2の実施例として、ヒット条件によって、ハイライト表示方法を変更する処理、さらに複数のハイライト処理を階層的に行なう場合の処理について説明する。概略処理ブロック図は、図1と同じである。図17は、本実施例で用いるヒット範囲情報(1701)である。図14に示したヒット範囲情報に対して追加された情報は、各ヒットした条件を格納する領域(1702)が追加されていることである。さらに、図14では、ヒットした検索タームの範囲だけを出力しているが、検索条件によって、ヒットした検索タームに加えて、その検索タームが含まれる構造全体に対するハイライトなど、検索タームを含む領域を指定することを可能としている。これらのヒット条件の情報は、構造化文書の検索処理時に付加する。ここでは、検索条件に用いられた距離条件、各検索タームの出現頻度などの情報を付加しているが、検索ターム毎にあらかじめ、重み付けを行なうなどの方法を用いることもできる。

【0043】図18は、ヒット条件とハイライト方法(ハイライト表示形態)の対応を定義したテーブル(1801)である。ヒット条件(1802)に対応するハイライト方法(1803)が記述されている。各ヒット条件によって、ヒットした範囲は、本テーブルの内容に基づいてハイライト表示を行なう。さらに、階層情報(1804)が与えられており、階層情報の値が大きいほど、構造全体のハイライトなど上位のハイライト構造となっている。

【0044】図19は、上記のハイライト処理を実現するための、表示用DTD作成の処理内容を示したものである。登録に用いた元のDTD(1901)に対して、上位のハイライト構造内には下位のハイライト構造を階層的に指定でき、さらに省略も可能なように定義を変更、追加したハイライト表示用のDTD(1902)を生成している。DTDの作成方法は、図11を用いて前述した処理に対して、1106ステップのハイライト情報付加の際に、複数存在するハイライト情報をすべて付加(1903)し、さらに1108ステップのハイライト用ELEMENT宣言追加の際に、図18の階層情報(1804)を元に、各ハイライト構造の下位構造として、下位のハイライト構造および文字列を内容モデルとして持つようにすれば良い。下位のハイライト構造がなけれ

ば、内容モデルとして、文字列だけが出現する（１９０４）。

【００４５】図２０は、第2の実施例におけるハイライト処理のフローチャートである。まず、ハイライト情報を開始位置順を第1キー、階層情報の上位から下位の順を第2キーとしてソートする（２００１）。次に、ハイライト開始までのテキストを出力し（２００２）、ハイライト開始タグを出力する（２００３）。さらに、ハイライト範囲の終わりまでに、次のハイライトが開始していれば、下位の構造情報が存在するため（２００４）、その位置までのテキストを出力した上で（２００５）、下位のハイライト構造におけるハイライト処理を行なう（２００６）。下位構造におけるハイライト処理は、２００３から２００９の処理と同じである。下位のハイライト構造に対する処理を終えた後、さらに下位のハイライト構造があれば（２００７）、２００５ステップの処理に戻って、次のハイライト構造までのテキストを出力し、下位のハイライト構造の処理を行なう。下位のハイライト構造がなくなれば、構造の終わりまでのテキストを出力して（２００８）、ハイライト終了タグを出力する（２００９）。ハイライトの情報が残っていれば、２００２のステップに戻り、処理を繰り返す。ハイライトの情報が終われば（２０１０）、残ったテキストを出力し、処理を終える（２０１１）。

【００４６】図２１は、上記処理により生成されるＳＧＭＬ文書の例である。図２２は、図２１のＳＧＭＬ文書の本文の表示例である。重なったハイライト範囲については、複数のハイライトのための表示方法を重複して行なっている。

【００４７】第3の実施例として、構造化文書の部分構造だけを切り出し、ハイライト表示する場合の処理内容を示す。図２３は、本実施例の概略処理ブロック図を示したものである。図1からの変更点は、表示対象の構造（２３０１）を指定するようにしていることと、表示対象の構造の指定内容を元に、文書表示（１１２）の処理の代わりに部分構造表示の処理（２３０２）を行なっていることである。

【００４８】図２４は、部分構造を抽出して、表示する場合の処理手順を示したフローチャートである。まず、部分構造表示用のＤＴＤを作成する（２４０１）。部分構造表示用のＤＴＤの作成処理については、図２６を用いて後述する。さらに、正規化後の構造化文書に対して得られたヒット範囲については、正規化前の登録時の文書における、構造ＩＤおよびヒット範囲情報に変換する（２４０２）。正規化後の文書の情報の正規化前の文書の範囲情報への変換処理の内容については、図１６を用いて前述した方法を用いることができる。次に表示対象となっている解析済み文書の構造の情報を読み出し、２４０３から２４１１の処理を順に繰り返すことで、表示用の文書の出力処理を行なう。まず、表示対象となる構造情報を読み出す（２４０３）。ここで表示対象の構造であるか否かの判定は、図7を用いて前述した方法を用いて実現する。表示対象の構造情報であれば、まず、構造の開始タグを出力する（２４０４）。さらに本構造に下位構造が存在するなら（２４０５）、下位構造に対して、表示処理（２４０３から２４１１の処理）を行なう（２４０６）。下位構造がなくなれば、構造の終わりを示すタグを出力する処理（２４１１）に移る。２４０５のステップで下位構造が存在しないと判定された場合は、文字列の構造であるため、本構造の内容に対して、ヒット範囲情報と比較し（２４０８）、ヒット範囲が含まれる構造であれば、ハイライト処理を行なう（２４０９）。ハイライト処理については、図１５を用いて前述した方法を用いる。ハイライト範囲が含まれない文字列であれば、内容をそのままテキストとして出力する（２４１０）。出力内容が文字列の場合は、２４０４、２４１１のステップで、開始タグ、終了タグは出力しない。上記の処理で構造ごとのハイライト表示を実現する。さらに処理すべき構造があれば、２４０３からの処理を繰り返す（２４１２）。

【００４９】図２５は、部分構造表示用のＤＴＤの作成内容である。部分構造の出力により、元のＤＴＤ（２５０１）で必ず出現しなければならないと定義されている構造が出力されない場合がある。さらに上位構造が必ずしも出力されるとは限らない。このため、部分構造表示用のＤＴＤは、上位構造の開始タグ、終了タグの出現を必須としない。さらに構造そのものについても、必ずしも出現しなくて良いとするように変更する必要がある。作成された部分構造表示用のＤＴＤは２５０２に示したようになる。このＤＴＤを用いて作成したＳＧＭＬ文書は、２５０３に示したようになる。この例では、タイトルだけを抽出している。

【００５０】図２６は、部分構造表示用のＤＴＤ作成手順を示したフローチャートである。まず、登録用のＤＴＤを取得する（２６０１）。次にＤＴＤ中のELEMENT項目を取り出す（２６０２）。内容モデルにCDATA、RCDATA、#PCDATAが含まれる場合は、ハイライト情報を付加する（２６０３－２６０６）。ハイライト情報の付加は、図１１の１１０３から１１０６ステップの処理と同じである。次に内容モデル中の出現指示子（*、+、?、なし）をチェックし、"+"ならば（２６０７）、"*"に変更し（２６０８）、出現指示子がなければ（２６０９）、"?"を付加する（２６１０）。全てのELEMENT宣言に対する処理が終わると（２６１１）、ハイライト用の構造のELEMENT宣言を追加し（２６１２）、さらに、下位構造が存在する構造のタグが出現することが必須(-)であれば、不要(O)に変更する。

【００５１】次に、本発明を用いた実施例4について、図面を用いて説明する。図２７は、本実施例のシステム構成図である。WWW（World Wide Web）検索シス

テム（２７００）は、ネットワーク（２７０２）を使用してクライアント（２７０１）と接続されている。クライアント（２７０１）は、ＰＣ，ＷＳなどであり、クライアント（２７０１）上で動作するＷｅｂブラウザ（２７０３）上の、検索ターム設定画面上で検索タームを入力する。ＷＷＷ検索システム（２７００）では、この検索タームを用いて検索を行い、その検索結果をＷｅｂブラウザ（２７０３）に出力する。ＷＷＷ検索システム（２７００）は、クライアント（２７０１）からの検索タームを受け取るＨＴＴＰサーバ（２７０４）と、検索処理およびハイライト用タグを挿入するデータ制御部（２７０５）と、ハイライトタグの位置情報などを格納しておくメモリ（２７０６）から成り立ち、検索対象となるＨＴＭＬ文書を格納しておく磁気ディスク装置（２７０７）が接続されている。データ制御部（２７０５）では、ＨＴＴＰサーバ（２７０４）で受け取った検索タームを磁気ディスク（２７０７）中に存在するＨＴＭＬ文書に対して検索処理を行い、検索タームにヒットしたＨＴＭＬ文書の検索ヒット位置にハイライトタグを挿入する。メモリ（２７０６）は、各文書ごとの検索ヒット数を格納するハイライト数格納領域（２７０８）と、検索結果位置情報を格納するハイライト位置情報格納領域（２７０９）と、挿入するハイライト用タグの内容を格納しておくハイライト用タグ文字格納領域（２７１０）と、ハイライト用タグを挿入したＨＴＭＬ文書を格納するＨＴＭＬ文書一時格納領域（２７１１）と、クライアント（２７０１）で入力した検索タームをＷＷＷ検索システム（２７００）のＨＴＴＰサーバ（２７０４）で取得し、一時的に格納する検索ターム格納領域（２７１２）からなる。ＷＷＷ検索システム（２７００）によってハイライト用タグを挿入したＨＴＭＬ文書は、ＨＴＴＰサーバ（２７０４）からネットワーク（２７０２）を経由して、クライアント（２７０１）のＷｅｂブラウザ（２７０３）上に表示される。

【００５２】次に、データ制御部（２７０５）の処理内容について、図２８を用いて説明する。ここでは、クライアント（２７０１）で設定した検索タームを取得し、検索処理を行い、検索ヒット位置を検出しハイライト位置情報（２７０９）を作成し、検索条件にヒットしたＨＴＭＬ文書の検索タームにヒットしたＨＴＭＬ文書の検索ヒット位置にハイライト用のタグを埋め込み、クライアント（２７０１）のＷｅｂブラウザ（２７０３）に表示する。

ステップ２８００：クライアント（２７０１）で設定した検索タームを、ＷＷＷ検索システム（２７００）では、ＨＴＴＰサーバ（２７０４）を用いて取得する。取得した検索タームは、メモリ（２７０６）の検索ターム格納領域（２７１２）に格納される。

ステップ２８０１：ステップ２８００で検索ターム格納領域（２７１２）に格納した検索タームを用いて、磁気ディスク装置（２７０７）に格納されているＨＴＭＬ文書に対する全文検索を行う。検索ヒットした場合は、ＨＴＭＬ文書中の検索ヒット位置や検索ヒット数などを取得し、その情報をハイライト位置情報格納領域（２７０９）、ハイライト数格納領域（２７０８）に格納する。この処理については、図２９を用いて詳しく説明する。

ステップ２８０２：ステップ２８０１において、作成されたハイライト位置情報格納領域（２７０９）に格納されている情報を基に、ハイライトタグ文字格納領域（２７１０）に格納されているハイライト用タグをＨＴＭＬ文書の検索ヒットした位置に挿入し、ＨＴＭＬ文書一時格納領域（２７１１）に格納する。詳細は、図３３を用いて説明する。

ステップ２８０３：ステップ２８０２により作成されたＨＴＭＬ文書一時格納領域（２７１１）に格納されたハイライト用ＨＴＭＬ文書を、ＨＴＴＰサーバ（２７０４）を用いてクライアント（２７０１）のＷｅｂブラウザ（２７０３）に表示する。ステップ２８００からステップ２８０３の処理を繰り返すことにより、クライアント（２７０１）で入力された検索条件を用いて、磁気ディスク（２７０７）に格納されているＨＴＭＬ文書を検索し、検索条件にヒットした文書に対して、複数箇所の検索ヒット位置のハイライト表示を可能とする。

【００５３】次に、図２９を用いて、図２８のステップ２８０１のハイライト位置情報の作成処理について説明する。

ステップ２９００：磁気ディスク（２７０７）に格納されているＨＴＭＬ文書を読み出す。図３４のＨＴＭＬ文書（３４００）は、読み出したＨＴＭＬ文書の例である。このＨＴＭＬ文書をＷｅｂブラウザで表示すると、３４０１に示すような画面が表示される。

ステップ２９０１：ハイライト位置情報を格納する領域であるハイライト位置情報格納領域（２７０９）をα件数分確保する。αは、任意の正の整数である。またハイライト数を格納するハイライト数格納領域（２７０８）を確保する。なお、ハイライト位置情報格納領域（２７０９）と、ハイライト数格納領域（２７０８）のデータ形式は、図３０および図３１に示す。ハイライト位置情報格納領域（２７０９）は、図３０に示すように、ＨＴＭＬ文書番号（３０００）、先頭からのハイライト位置番号（３００１）、ハイライトバイト数（３００２）、ハイライト挿入タグ番号（３００３）から構成される。ＨＴＭＬ文書番号（３０００）は、ステップ２９００で読み出したＨＴＭＬ文書の番号である。ＨＴＭＬ文書を格納した際に付けられる通し番号などを格納する。先頭からのハイライト位置番号（３００１）は、ステップ２９００で読み出したＨＴＭＬ文書にステップ２８００で取得した検索タームにヒットした場合、ＨＴＭＬ文書中の検索ヒット位置を文書先頭からバイト数で格納する。ハイライトバイト数（３００２）は、ハイライトする長

さをバイト 数で格納する。つまり 、検索タームの文字列長を格納する。ハイライト 挿入タグ番号(３００３) は、複数の検索タームでハイライト 表示する場合、検索タームごとにハイライト 用タグを区別して表示することが可能である。ここに格納されている情報を基にして、ハイライト 用タグを区別する。つまり 、ここには、ハイライト 表示に利用するタグの種類を判別するデータを格納する。

【００５４】ステップ２９０２： ハイライト 位置情報格納領域(２７０９) に格納したカウントを示すi_cntを０に初期設定する。

ステップ２９０３： ステップ２８００で読み出した検索タームとステップ２９００で読み出したHTML 文書が一致するか否かをチェックをする。検索ヒット 箇所が存在する場合は、ステップ２９０４に進む。また、存在しない場合は、ステップ２９０８に進む。

ステップ２９０４： ステップ２９０１または２９０５で確保したハイライト 位置情報格納領域(２７０９) がハイライト 格納数を示すi_cntより 大きいか否かをチェックする。データを格納する領域がまだ存在する場合、ステップ２９０６に進む。また、格納する領域が存在しない場合、ステップ２９０５に進む。

ステップ２９０５： ハイライト 位置情報格納領域(２７０９) を一定値拡大して再度確保し直し、ステップ２９０６に進む。

【００５５】ステップ２９０６： ステップ２９０１または２９０５で確保したハイライト 位置情報格納領域(２７０９) のi_cnt番目の位置に、HTML 文書番号(３０００) 、HTML 文書の先頭からの位置(３００１) 、ハイライト 文字数(３００２) 、ハイライトタグ挿入番号(３００３) を格納する。i_cntは0 に初期化されているので、i_cntが0 の場合、０番目にデータを格納する。１つのHTML 文書中に複数のハイライト 情報を格納する場合は、i_cntが更新されるので、i_cntが示す位置に格納する。ステップ２９００で読み出したHTML 文書(３４００)をHTML 文書番号「００１」とする。さらに、ステップ２８００で抽出した検索タームを「 特集」とする。このHTML 文書(３４００) で、検索ターム「 特集」を検索すると、HTML 文書(３４００) の先頭から１２２バイト 目(３４０３) に「 特集」の文字を見つけることができる。この場合、HTML 文書番号(３０００) にはHTML 文書番号である「００１」(３４０４) を格納し、HTML 文書の先頭からの位置(３００１) には「１２２」(３４０５) を格納し、ハイライト 文字数(３００２) には「 特集」のバイト 数「４」(３４０６) を格納する。最後に、ハイライトタグ挿入番号(３００３) には、検索結果を強調するためのタグを示す番号を格納する。ここでは、「１」(３４０７) を格納する。

【００５６】ここで、ハイライト 挿入タグ番号と 実際に

格納するハイライトタグを対応する構成を図３２に示す。図３２の(１) では、ハイライトタグ文字格納領域(２７１０) に格納されているハイライト 挿入タグ用の構造体３２００を示す。ハイライト 挿入タグ用の構造体(３２００) は、通し番号を格納するタグ番号1 (３２０２) と、ハイライト 開始タグ名を格納する開始タグ１(３２０３) 、ハイライト 終了タグ名を格納する終了タグ１(３２０４) と、タグの個数を格納するハイライトタグ数(３２０１) から成り 立つ。ハイライトタグ数に格納した数分のタグ番号、開始タグ、終了タグが存在する。

【００５７】ハイライトタグ文字格納領域の使用例を(２) に説明する。ここでは、３種類のハイライトタグを格納する場合を示す。よって、ハイライトタグ数を格納する箇所には、「３」(３２０５) を格納する。タグ番号「０」(３２０６) の箇所の開始タグには、赤色を示すタグ「 <FONT COLOR＝“RED”>」(３２０７) を、終了タグには「 </FONT>」(３２０８) を格納する。同様に、タグ番号「１」(３２０９) には、点滅を示すタグ「 <BLINK>」を、タグ番号「２」(３２１０) には、文字を大きく表示する「 <H1 >」を格納する。ハイライトタグ文字格納領域(２７１０) は、ハイライト 位置情報格納領域(２７０９) の作成前に作成する。また、このハイライトタグ文字格納領域(２７１０) は、ユーザインターフェースを使用して、作成することも 可能である。複数のハイライト 用タグを用意することで、異表記や同義語の検索処理を行った場合において、異表記で検索された文字にはタグ番号「１」、同義語で検索された文字にはタグ番号「２」のように、検索条件ごとに異なるハイライト 表示が可能となる。ハイライト 用タグに「 <BLINK>」を使用する場合は、ハイライト 位置情報格納領域(３４０２) のハイライトタグ挿入番号(３４０７) に「１」を格納する。

【００５８】ステップ２９０７： ステップ２９０６において、ハイライト 位置情報格納領域(２７０９) にデータを格納したので、i_cntを1 を加え、ステップ２９０３に戻る。

ステップ２９０８： ステップ２９００で取得したHTML 文書中のハイライト 数を取得し、ハイライト 数格納領域(２７０８) に格納する。ハイライト 数格納領域(２７０８) の構造体の内容は図３１を用いて説明する。図３１は、ハイライト 数格納領域(２７０８) の構造体の内容である。３１００は、ステップ２９００で読み出したHTML 文書の文書番号である。また、３１０１は、取得したハイライト 数を格納しておく 箇所である。ここでは、文書番号「００１」を文書番号３１００に格納し、i_cntをハイライト 数格納領域(３１０１) に格納し、処理を終了する。

【００５９】次に、図３３を用いて、ハイライト 用タグ

付のＨＴＭＬ文書作成処理について説明する。

ステップ３３００：ステップ２９００で読み出したＨＴＭＬ文書中に、ハイライトタグを挿入する必要があるか否かをチェックする。ハイライト位置情報格納領域（２７０９）に格納したＨＴＭＬ文書番号（３０００）が存在する場合は、ステップ３３０１に進む。存在しない場合は、ステップ３３０９ですべてのテキストを出力し、処理を終了する。

ステップ３３０１：処理カウントを示すi_cntを0に初期化する。

ステップ３３０２：ハイライトタグを挿入したＨＴＭＬ文書を格納するＨＴＭＬ文書一時格納領域（２７１１）を確保する。ＨＴＭＬ文書一時格納領域（２７１１）は、ＨＴＭＬ原文書のバイト数は、ハイライト用開始タグと終了タグのバイト数の合計値にハイライト挿入数を乗じたバイト数の領域を確保する。ハイライトの開始タグと終了タグは、ハイライト位置情報格納領域（２７０９）のハイライト挿入タグ番号（３００３）より、ハイライト用タグのタグ文字列長を計算する。ハイライト数は、ステップ２９０８でハイライト数格納領域（２７０８）に格納したハイライト数（３１０１）を取得する。

ステップ３３０３：ハイライト数（３１０１）がi_cntより小さいか否かをチェックする。小さい場合は、未処理のハイライト箇所が存在するので、ステップ３３０４に進む。それ以外は、処理すべき未処理のハイライト箇所を全て終了したので、ステップ３３０９に進む。

ステップ３３０４：ハイライト位置までのＨＴＭＬ文書をステップ３３０２で確保したＨＴＭＬ文書一時格納領域（２７１１）に格納する。

【００６０】ステップ３３０５：ハイライト開始タグをＨＴＭＬ文書一時格納領域（２７１１）に格納する。ハイライト開始タグは、ハイライト挿入タグ番号（３００３）から抽出した番号より得られるハイライトタグ文字格納領域（２７１０）に格納されているタグ名である。図３４（３）の場合、ハイライト挿入タグ番号（３００３）には「１」が格納されている。図３２（２）に示したハイライトタグ文字格納領域（２７１０）のタグ番号「１」（３２０９）に格納されている「＜ＢＬＩＮＫ＞」をＨＴＭＬ文書一時格納領域（２７１１）に格納する。

ステップ３３０６：検索タームをＨＴＭＬ文書一時格納領域（２７１１）に格納する。図３４の場合、「特集」をＨＴＭＬ文書一時格納領域（２７１１）に格納する。

ステップ３３０７：ハイライト終了タグをＨＴＭＬ文書一時格納領域（２７１１）に格納する。ハイライト終了タグは、ステップ３３０５で処理したハイライト開始タグ同様、ハイライト挿入タグ番号（３００３）にて格納された番号から得られるハイライトタグ文字格納領域（２７１０）に格納されているタグ名を格納する。図３４（３）の場合、「１」が格納されている。したがって、図３２（２）のタグ番号「１」に格納されている「＜／ＢＬＩＮＫ＞」をＨＴＭＬ文書一時格納領域（２７１１）に格納する。

ステップ３３０８：ステップ３３０５からステップ３３０７において、データをＨＴＭＬ文書一時格納領域（２７１１）に格納した後、i_cntに1を加え、ステップ３３０３に戻る。

ステップ３３０９：ハイライト挿入位置からＨＴＭＬ文書最後までテキストをＨＴＭＬ文書一時格納領域（２７１１）に格納し、ハイライトタグ付きＨＴＭＬ文書の作成処理を終了する。

【００６１】以上の処理を用いることで、クライアント（２７０１）設定した検索タームから、ＨＴＭＬ文書を検索し、検索タームと一致する文書に対して、ハイライト数を格納するハイライト数格納領域（２７０８）、ハイライト位置を格納するハイライト位置情報格納領域（２７０９）の内容を作成することが可能である。上記の処理結果の例を図３５に示す。３５００は、ハイライト用タグを挿入したＨＴＭＬ文書である。検索ヒットした「特集」の前後（３５０１，３５０２）にハイライト用タグが挿入されている。このＨＴＭＬ文書を画面に表示すると３５０３のようになり、検索ヒットした「特集」（３５０４）が点滅表示される。以上で、本発明の第1実施例として、クライアント（２７０１）が挿入した検索タームをＨＴＭＬ文書（２７０７）中から検索し、ヒット位置にハイライト用タグを挿入した、ハイライト用タグ付きＨＴＭＬ文書を作成する方法を説明した。

【００６２】次に、本発明における実施例5について、図３６から図４２を用いて説明する。図３６は、検索タームがＨＴＭＬ文書のタグで分断されている場合や、検索タームがタグ内に存在する場合のハイライト表示方法におけるシステム構成図である。図２７と同様に、クライアント（２７０１）のWebブラウザ（２７０３）上で検索タームを設定する。

【００６３】WWW検索システム（２７００）は、検索タームを取得するＨＴＴＰサーバ（２７０４）、検索処理を行うデータの制御（２７０５）、領域を確保するメモリ（２７０６）から成り立つ。メモリ（２７０６）は、図２７の説明で述べた以外に、レイアウト表示などに使用されるタグで、読み飛ばすタグ名を格納した読み飛ばしタグ名格納領域（３６００）と、クライアント（２７０１）が入力した検索タームとＨＴＭＬ文書（２７０７）が一致した開始位置を一時的に格納しておく開始位置格納領域（３６０１）と、検索タームとヒットした位置がＨＴＭＬタグの開始文字「＜」と終了文字「＞」の間に存在した場合、目印となるマークを格納しておく再度記述マーク格納領域（３６０２）と、ＨＴＭＬタグの開始タグと終了タグの間に検索タームがヒットした場合、検索ヒットした箇所の前後にハイライト用タ

グを入れることができないHTMLタグを記述しておく、ハイライトタグ挿入不可能タグ名格納領域（３６０３）からなる。検索タームがHTML文書のタグをまたがっている場合や、検索タームがタグ内に存在する場合の検索タームの取得、ハイライト位置情報の作成、ハイライト用タグ挿入方法は、図２８で示した処理手順で行う。また、各々の処理内容については、図３７から図４２を用いて説明する。

【００６４】ステップ２８００で取得した検索タームを用いて、ステップ２８０１の処理では、検索処理およびハイライト位置情報作成処理を行う。処理内容は、図３７のフローチャートに示す。

ステップ３７００：処理対象となるHTML文書を磁気ディスク（２７０７）から読みだす。

ステップ３７０１：ハイライト位置情報を格納するハイライト位置情報格納領域（２７０９）とハイライト数格納領域（２７０８）をメモリ（２７０６）に確保する。

ステップ３７０２：検索ヒット位置の前後に挿入するハイライトタグを読み出す。図３２（２）の使用例に具体例を示したようにハイライトタグ文字格納領域（２７１０）からハイライト用タグを読み出す。この場合ハイライト挿入タグ番号の個数は、「３」（３２０５）から「３つ」とわかる。１番目の「０」（３２０６）には、「＜FONT　COLOR＝“RED”＞」（３２０７）と「＜／FONT＞」（３２０８）格納されている。そこで、ハイライト挿入タグ番号0番目の開始タグは「＜FONT　COLOR＝“RED”＞」、終了タグは「＜／FONT＞」となる。同様に、ハイライト挿入タグ番号1番目の開始タグは「＜BLINK＞」、終了タグは、「＜／BLINK＞」となり、ハイライト挿入タグ番号2番目の開始タグは「＜H1＞」、終了タグは「＜／H1＞」となる。

ステップ３７０３：HTML文書の処理済み文字数のカウントを示すi_cntと、ハイライト数を格納する領域の内容を0に初期設定する。

【００６５】ステップ３７０４：検索タームとHTML文書の文字列が一致するか否かをチェックする。チェック方法として、HTML文書のi_cntバイト目から、検索タームの先頭文字と一致する文字を検索する。ステップ３７０３において、初期設定が0に設定されているため、最初は、HTML文書の0バイト目から一致する文字を検索する。一致した場合は、ステップ３７０５に進む。不一致の場合は、処理を終了する。また、ここでは、検索タームを抽出する方法として、指定したタグを飛ばして検索する方法を用いる。具体的には、読み飛ばしタグ名格納領域（３６００）に格納してあるタグ名をHTML文書中に存在した場合は、そのタグを読み飛ばし、検索処理を行う。読み飛ばしタグ名格納領域（３６００）に「ＩMG」を格納しておき、図３４のHTML文書（３４００）を検索した場合、HTML文書（３４００）中の先頭からデータを走査し、「ＩMG」（３４１３）が抽出された時点で、タグ内の文字を飛ばす。つまり、タグの終了文字「＞」（３４１４）までを飛ばす。この読み飛ばしタグ名格納領域（３６００）は、検索処理前に設定しておくことにより、読み飛ばし処理が可能となる。

【００６６】ステップ３７０５：ステップ３７０４でHTML文書の先頭から検索タームの先頭文字と一致した文字までの文字数を開始位置格納領域（３６０１）に一時的に確保する。

ステップ３７０６：検索タームの文字列とHTML文書に書かれている文字が一致するか否かをチェックし、一致した場合、一致箇所がHTMLタグ内に存在するかあるいはHTMLタグ外に存在するか否かをチェックする。さらに、検索ヒットした文字列の最後の文字の位置を、HTML文書の先頭からの文字数で確保する。詳細は、図３８を用いて説明する。

ステップ３７０７：ステップ３７０６の結果、検索ヒットしたか否かチェックする。HTML文書中に検索タームが存在した場合は、ステップ３７０８に進む。検索タームが存在しない場合、ステップ３７１２に進む。

ステップ３７０８：ステップ３７０１で確保したハイライト数格納領域（３７０８）とハイライト格納数を比較して、確保した領域が格納したハイライト数より多ければ、ステップ３７０９に進む。少なければ、ステップ３７１０に進む。

ステップ３７０９：ハイライト位置情報格納領域（２７０９）にデータを格納する領域が足りないため、再度領域設定し直し、ステップ３７１０に進む。

【００６７】ステップ３７１０：ハイライトする文字数とハイライトの位置の情報を、ハイライト位置情報格納領域（３６００）に格納する。具体的には、図３０で説明したハイライト位置情報格納領域（３６００）のHTML文書番号（３０００）には、ステップ３７００で読み出したHTML文書の文書番号を格納し、先頭からのハイライト位置情報（３００１）には、ステップ３７０５で取得した開始位置を格納する。また、ハイライトのバイト数（３００２）には、検索タームの文字列長を格納し、ハイライト挿入タグ番号（３００３）には、ステップ３７０２で読み出したタグの番号を格納する。ハイライト挿入タグ番号（３００３）は、デフォルトとして、「０」を設定する。

ステップ３７１１：検索タームにヒットする文字列が複数存在する場合、検索ヒットした位置の次文字から再度検索タームとHTML文書中の一致する箇所をチェックする処理を行う。そこで、ステップ３７０６で確保した検索ヒットの最後の文字が記述されている位置の、HTML文書の先頭からの文字数に1を加えた値をi_cntに代入する。処理位置を更新したら、ステップ３７０４に戻る。

29

ステップ３７１２：ステップ３７０５で取得した開始位置格納領域（３６００）に格納してある開始位置からの文字列と、検索タームが一致していない場合、開始位置の次文字から再度検索タームとHTML文書中の一致する箇所をチェックする処理を行う。そこで開始位置格納領域（３６００）に格納してある開始位置に1を加えた値をi_cntに代入する。処理位置を更新したら、ステップ３７０４に戻る。以上で、タグ内およびタグ外のチェックを含む検索処理およびハイライト位置情報作成処理について述べた。

【００６８】次に、図３８を用いて、ステップ３７０６のタグ内の検索およびタグ外の検索処理について説明する。ここでは、ステップ３７０５で取得した検索ヒットの開始位置が、文書構造を示すタグの属性中に存在するかあるいはタグの外に存在するかをチェックし、検索ヒットの開始位置からの文字列が検索タームと一致するか否かのチェックを行う。

ステップ３８００：ステップ３７０５で開始位置格納領域（３６００）に格納した検索ヒットの開始位置において、HTMLタグ内かあるいはタグ外かをチェックする。ステップ３７０６時点におけるHTML文書のi_cntバイト目から、検索ヒットの開始位置までのデータをチェックする。タグの開始文字「＜」と対応するタグの終了文字「＞」をチェックし、タグ内に検索ヒットの開始位置が存在するか否かをチェックする。タグの開始文字「＜」があり、タグの終了文字「＞」の前に検索ヒットの開始位置が存在する場合、開始位置はタグ内に存在するとして、ステップ３８０１に進む。タグの開始文字「＜」とタグの終了文字「＞」に囲まれない範囲に、検索ヒットの開始位置が存在する場合、検索ヒットの開始位置は、タグ外に存在するとして、ステップ３８０４に進む。

【００６９】ステップ３８０１：検索タームと、検索ヒットの開始位置からの文字列が一致するか否かをチェックする。検索タームの文字列が複数バイトから成り立つ場合、文字列を1バイト毎にチェックする。検索タームの文字列と検索ヒットした位置からの文字列が一致する場合、ステップ３８０２に進む。不一致の場合、ステップ３８０３に進む。

ステップ３８０２：ステップ３８０１において、検索タームと一致した場合、「検索ヒット」したとして、処理を終了する。また、検索ヒットした文字列の終端位置を求める。終端位置は、検索ヒットした開始文字位置に検索タームの文字列長を加えたバイト数とする。ここで求めた終端位置は、ステップ３７１１にて使用される。

ステップ３８０３：ステップ３８０１において、検索タームが不一致の場合、「検索ヒットしない」として、処理を終了する。

ステップ３８０４：ステップ３８００において、検索ヒットの開始位置がタグの外に存在した場合、タグ外用の検索処理を行う。タグ外用の検索処理は、図３９を用いて説明する。

30

ステップ３８０５：ステップ３８０４で検索タームがヒットする箇所がHTML文書中に存在するか否かをチェックする。存在する場合は、ステップ３８０７に進む。存在しない場合は、ステップ３８０６に進む。

ステップ３８０６：ステップ３８０５において、検索タームがヒットしない場合、処理を終了する。

ステップ３８０７：ステップ３８０５において、検索タームと一致した場合、「検索ヒット」したとして、処理を終了する。また、検索ヒットした文字列の終端位置を求める。終端位置は、検索ヒットの開始文字位置に、ステップ３８０４で検出した検索ヒットの最後の文字が記述されている位置を加えた値とする。ここで求めた終端位置は、ステップ３７１１にて使用する。以上で、タグ内検索およびタグ外検索処理について説明した。

【００７０】次に、ステップ３８０４のタグ外用の検索処理について、図３９を用いて説明する。

ステップ３９００：HTML文書中に検索タームが存在するか否かをチェックする。検索タームの文字列がHTML文書中に存在する文字列と一致するか否かをチェックするが、検索ヒットした開始位置から、途中に存在するタグを飛ばすことにより一致する場合があるので、ここでは、検索ヒットした開始位置から1文字ごとに検索タームと合致しているか否かをチェックする。具体的には、図３４を用いて説明する。検索タームを「特集記事」とした場合、（２）の表示画面では、３４０８に「特集記事」が表示されている。しかし、HTML文書（３４００）では、「特集」（３４０３）と「記事」（３４１６）の間に「＜／H1＞」（３４１７）のタグがある。このように検索タームの途中にHTMLタグが存在する場合、このHTMLタグを読み飛ばして、検索タームと一致する文字列を抽出する。ここでは、検索タームを1文字ごとHTML文書の文字と照合し、チェックを行う。検索タームの1文字目とHTML文書中の文字が一致した場合は、検索タームの次文字とHTML文書の次文字について、同処理を繰り返す。検索タームの文字列のすべての文字が一致した場合、具体的には、「特」（３４０３）、「集」、と文字の比較を行い、「＜／H1＞」（３４１７）を読み飛ばし、さらに、「記」（３４１６）、「事」と文字比較を行う。すべての検索タームを抽出し終わった場合、ステップ３９０１に進む。検索タームがHTML文書中の文字列と完全に一致しなかった場合、ステップ３９０２に進む。

【００７１】ステップ３９０１：HTML文書中に検索タームが存在するため、「検索ヒット」したとして、処理を終了する。また、検索ヒットの終端位置を求める。終端位置は、ステップ３９００において、最後に抽出した文字の位置である。

ステップ３９０２：ステップ３９００で、検索タームの

31

文字とHTML 文書の文字が一致しなかった場合、HTML 文書の文字が、タグの開始文字「 <」か否かをチェックする。タグの開始文字「 <」の場合は、ステップ3903に進む。それ以外の文字の場合は、ステップ3904に進む。

ステップ3903：ステップ3902において、HTML 文書中の文字がタグの開始文字「 <」の場合、タグの内容を飛ばして、ステップ3900に戻る。具体的には、タグの終了文字「 >」を抽出し、抽出した文字までを読み飛ばす。図34のHTML 文書（3400）で、検索タームを「 特集記事」とした場合、「 特集」（3403）の次文字にある「 <」（3417）からタグの終了文字「 >」（3418）までを読み飛ばす。つまり、「 </H1 >」を読み飛ばす。

ステップ3904：ステップ3902において、検索タームが不一致の場合、「 検索ヒットしない」として、処理を終了する。以上で、HTML 文書中に検索タームの文字列が存在した場合、検索ヒット 位置を抽出し、ハイライト 位置情報格納領域にハイライト 位置情報を格納する処理について説明した。

【0072】次に、図40を用いて、ハイライト 位置情報格納領域に格納したハイライト 位置情報を基にして、HTML 文書の検索ヒットした文字列を強調表示するため、ハイライト 用タグを挿入する方法について説明する。

ステップ4000：ステップ3710においてハイライト 情報格納領域（2709）に格納した、ハイライト 位置情報を読み出す。

ステップ4001：ハイライトタグを挿入したHTML 文書を格納するためのHTML 文書一時格納領域（2711）を確保する。確保する領域の大きさは、HTML 文書の原文書のデータ、ハイライトタグ数分のハイライト 開始タグと終了タグの長さの和を乗じた値のバイト 数である。ハイライトタグ数は、ハイライトタグ数格納領域（2708）から読み出す。また、ハイライト の開始タグと終了タグは、ハイライト 位置情報格納領域（2709）のハイライト 挿入タグ番号（3003）とハイライトタグ文字格納領域（2710）からタグを検出し、検出したタグの文字列長を求める。

ステップ4002：HTML 文書中の処理済み位置を示すi_cntと、ハイライト 処理数を0 に初期化する。

【0073】ステップ4003：処理済みのハイライト 箇所の数であるハイライト 処理数と、処理すべきハイライト 数を比較する。ハイライト 処理数が少ない場合は、ハイライト 用タグを挿入する処理を行うため、ステップ4004に進む。それ以外の場合は、ステップ4007に進む。

ステップ4004：処理済みの位置を示すi_cntから検索ヒット の開始位置までのデータを、HTML 文書一時格納領域（2711）に格納する。具体的には、図34

32

のHTML 文書（3400）で、検索タームを「 特集記事」とした場合、HTML 文書先頭から「 特集記事」（3403）前の文字「 今月の」までのデータをHTML 文書一時格納領域（2711）に格納する。

ステップ4005：ハイライト 用タグを検索ヒット 位置に格納する。ハイライト 用タグの挿入処理については、図41で説明する。

ステップ4006：HTML 文書の処理済み位置を示すi_cntにハイライト 終了タグを挿入した位置の先頭からのバイト 数を代入し、ステップ4003に戻る。

ステップ4007：HTML 文書の処理済み位置を示すi_cntから、HTML 文書の最後までのデータをHTML 文書一時格納領域（2711）に格納し、処理を終了する。

【0074】次にステップ4005で処理するハイライトタグの挿入処理について、図41を用いて説明する。ここでは、検索ヒットした位置が、タグの内あるいはタグ外かをチェックし、検索ヒット 位置の前後にハイライト 用タグを挿入する処理を行う。

ステップ4100：HTML 文書でヒットした位置がHTML タグ内か、タグ外かをチェックする。チェック方法は、検索ヒット の開始位置までのHTML 文書において、HTML タグの開始文字「 <」とタグの終了文字「 >」の対応をとり、タグ内かタグ外かを判断する。検索ヒット の開始位置がタグの開始文字「 <」からタグの終了文字「 >」の間にある場合は、タグ内に存在するとして、ステップ4101に進む。それ以外の場合は、タグ外に検索ヒット 位置の開始位置が存在するとし、ステップ4110に進む。

ステップ4101：タグの開始文字「 <」の次文字から文字を抽出し、タグの種類を取得する。例えば、図34のHTML 文書（3400）の場合、検索タームを「 hitachi」とした場合、HTML 文書（3400）中の3409に「 hitachi」を取得することができる。このHTML タグの種類を取得すると、タグの開始文字「 <」の次に書かれている「 A」（3410）とわかる。

ステップ4102：ステップ4101で取得したタグが、開始用のタグが終了用のタグかをチェックする。終了用のタグの場合、タグの開始文字「 <」の次文字が「 /（スラッシュ）」である。そこで、タグの開始文字「 <」の次文字をチェックし、判別する。このタグ開始文字「 <」の次文字が「 /」の場合、終了用のタグと判定して、ステップ4105に進む。それ以外の場合は、開始タグと判定し、ステップ4103に進む。

【0075】ステップ4103：開始用タグと終了用タグの間にハイライト 用のタグを挿入することが可能か否かをチェックする。挿入することが可能な場合は、ステップ4105に進む。また、不可能な場合は、ステップ4106に進む。具体的には、図34のHTML 文書（3400）で、検索タームが「 hitachi」の場合、H

33

ＴＭＬタグの種類は、「 hitachi」（３４０９）の前に存在する「Ａ」（３４１０）である。開始用のタグとは、ＨＴＭＬタグの種類を囲む「＜」と「＞」に挟まれた部分の「＜Ａ～＞」（３４１０）であり、終了用のタグとは、「＜／Ａ＞」（３４１２）である。このＨＴＭＬタグは、開始用タグと終了用タグに書かれた文字列が画面に表示され、この文字列を選択すると、開始用タグ内で指定したＵＲＬにリンクするタグである。この「＜Ａ～＞」（３４１０）と「＜／Ａ＞」（３４１２）の間にハイライト用のタグを挿入すると、強調表示が正しく行われない現象が生じる。そこで「＜Ａ～＞」タグと「＜／Ａ＞」タグの間にはハイライト用のタグを挿入することができないとして、ステップ４１０４に進む。ハイライトタグの挿入の可能あるいは不可能の判定は、ＨＴＭＬタグの種類がハイライトタグ挿入不可能タグ名格納領域（３６０３）に格納されているタグのいずれかと一致すれば、不可能と判定し、いずれとも不一致の場合は、挿入可能と判定する。なお、ハイライトタグ挿入不可能タグ名格納領域（３６０３）に格納するタグは、ユーザインタフェースを使用して、データ制御（２７０５）の処理を行う前に、作成しておく。

【００７６】ステップ４１０４：開始用タグと終了用タグの間に検索ヒットした文字列を挿入することが出来ない場合、挿入可能な箇所まで読み飛ばす。ここでは、終了用タグの終わりの文字「＞」まで読み飛ばし、読み飛ばしたデータを取得する。具体的には、図３４のＨＴＭＬ文書（３４００）、検索タームを「 hitachi」とした場合、ＨＴＭＬタグは、「Ａ」（３４１０）である。その終了タグは「＜／Ａ＞」（３４１２）である。よって「＜／Ａ＞」（３４１２）までデータを取得する。

ステップ４１０５：開始用タグと終了用タグの間に文字挿入することができる場合、タグの終わりの文字「＞」までデータを取得する。具体的には、図３４のＨＴＭＬ文書（３４００）検索タームが「 imagefile.gif」（３４１１）とした場合、ＨＴＭＬタグの種類は、「ＩＭＧ」（３４１３）で、その終わりのタグ文字は、「＞」（３４１４）となる。よって、「＞」（３４１４）まで読み飛ばし、読み飛ばしたデータを取得する。

ステップ４１０６：ステップ４１０４、ステップ４１０５で取得したデータをＨＴＭＬ文書一時格納領域（２７１１）に挿入する。また、検索ヒットした文字の終端位置を求める。終端位置は、ステップ４１０４あるいはステップ４１０５で読み飛ばした位置のバイト数である。

【００７７】ステップ４１０７：ハイライト用の開始タグをＨＴＭＬ文書一時格納領域（２７１１）に挿入する。図３２の開始タグ１（３２０３）に書かれているタグ名をＨＴＭＬ文書一時格納領域（２７１１）に挿入する。ハイライト位置情報格納領域に格納されているデータを３４０２とし、ハイライトタグ文字格納領域を図３２の（２）とした場合、ハイライト位置情報格納領域の

34

ハイライト挿入タグ番号（３４０７）に「１」が記述してあるので、図３２の（２）から、「＜ＢＬＩＮＫ＞」を取得することができる。よって、ここでは、「＜ＢＬＩＮＫ＞」をＨＴＭＬ文書一時格納領域（２７１１）に挿入する。

ステップ４１０８：検索ヒット文字をＨＴＭＬ文書一時格納領域（２７１１）に挿入する。但し、再度検索ヒット文字を記述するため、再度記述したマークを格納する。ここでは、再度記述マークの格納領域（３６０２）に格納されているデータを挿入する。再度記述マークの格納領域（３６０２）は、データ制御を行う前に、データを格納しておく。例えば、「＠」マークを格納しておく。

ステップ４１０９：ハイライト用の終了タグをＨＴＭＬ文書一時格納領域（２７１１）に挿入する。図３２の終了タグ１（３２０４）に書かれているタグ名をＨＴＭＬ文書一時格納領域（２７１１）に挿入する。ここでは、「＜／ＢＬＩＮＫ＞」を挿入する。

ステップ４１１０：検索ヒットした位置がタグの外に存在する場合、タグ外用ハイライトタグ挿入処理を行う。この処理については、図４２を用いて説明する。次にステップ４１１０で処理をするタグ外用ハイライト挿入処理について、図４２を用いて説明する。

ステップ４２００：「＜Ａ～＞」タグの場合、開始用のタグ「＜Ａ～＞」と終了用のタグ「＜／Ａ＞」に挟まれている箇所にハイライト用のタグを挿入すると、強調表示が反映されない。ここでは、開始タグと終了タグに囲まれているか否かをチェックし、ステップ４２０１でハイライトタグを挿入することが可能なタグか否かを判断する。但し、ＨＴＭＬ文書には、「＜ＨＴＭＬ＞」と「＜ＢＯＤＹ＞」のタグが存在するが、これらのタグは、文書全体を囲むタグとして使用されているため、ここでは、チェックの対象としない。図３０の先頭からのハイライト位置情報（３００１）に格納されている位置が、ＨＴＭＬ文書の開始用タグと終了用タグに囲まれているか否かをチェックする。囲まれている場合は、ステップ４２０１に進む。囲まれていない場合は、ステップ４２０９に進む。具体的には、図３４のＨＴＭＬ文書（３４００）で、検索タームが「日立」とした場合、ＨＴＭＬ文書（３４００）の「日立」（３４１５）が抽出される。これは、ＨＴＭＬタグ「＜Ａ～＞」（３４１０）と「＜／Ａ＞」（３４１２）に囲まれているので、ここでは、ステップ４２０１に進む。

ステップ４２０１：検索ヒットした文字列の前後にハイライト用タグを挿入することが可能か否かをチェックする。チェック方法は、検索ヒットした検索文字列に囲まれているＨＴＭＬタグを抽出する。抽出したタグの種類とハイライトタグ挿入不可能タグ名格納領域（３６０３）に格納されているタグと比較する。一致すれば、ステップ４２０２に進み、不一致の場合、ステップ４２０

9に進む。ハイライトタグ挿入不可能タグ名格納領域（３６０３）中に記述されたタグ名は、開始用タグと終了用タグの間に、ハイライト用タグを挿入することができない。よって、ハイライトタグ挿入不可能タグ名格納領域（３６０３）に格納されているＨＴＭＬタグと比較し、一致すれば、ステップ４２０２に進む、不一致の場合は、ステップ４２０９に進む。このハイライトタグ挿入不可能タグ名格納領域（３６０３）は、ユーザインタフェースを使用して、データ制御（２７０５）の前に作成しておく。

ステップ４２０２：ハイライト用タグを挿入することが出来ない場合、終了用タグの終わりの文字「＞」までのＨＴＭＬ文書を読み飛ばす。図３４において、検索ターム「日立」とした場合、「＜／Ａ＞」（３４１２）までのＨＴＭＬ文書を読み飛ばす。

【００７８】ステップ４２０３：ステップ４２０２で飛ばしたＨＴＭＬ文書をＨＴＭＬ文書一時格納領域（２７１１）に格納する。図３４において、検索ターム「日立」とした場合、ステップ４００２で設定したi_cnt番目あるいはステップ４００６で更新したi_cnt番目にあるＨＴＭＬ文書の文字から「＜／Ａ＞」（３４１２）のデータをＨＴＭＬ文書一時格納領域（２７１１）に格納する。

ステップ４２０４：ハイライト用タグの開始タグをＨＴＭＬ文書一時格納領域（２７１１）に挿入する。ハイライト位置情報格納領域を３４０２として、ハイライトタグ文字格納領域を図３２の（２）とした場合、「＜ＢＬＩＮＫ＞」が抽出される。よって、ここでは、「＜ＢＬＩＮＫ＞」を挿入する。

ステップ４２０５：再表示用マークを格納する。ステップ４１０８同様に、再度記述マークの格納領域（３６０２）に格納されているＨＴＭＬ文書を読み出し、ＨＴＭＬ文書一時格納領域（２７１１）に格納する。

ステップ４２０６：検索ヒットした文字列をもう一度ＨＴＭＬ文書一時格納領域（２７１１）に挿入する。但し、検索ヒットした文字列の途中にタグが存在する場合は、タグが存在する箇所までの文字列を挿入する。

ステップ４２０７：ハイライト用の終了タグをＨＴＭＬ文書一時格納領域（２７１１）に挿入する。ここでは、「＜／ＢＬＩＮＫ＞」を挿入する。

ステップ４２０８：ステップ４２０６において、検索タームの文字列をすべて格納したか否かをチェックする。検索ヒットした文字列中にタグが存在し、検索ヒットした文字をすべて格納していない場合、ステップ４２００に戻る。また、すべての文字を格納した場合は、処理を終了する。

【００７９】ステップ４２０９：検索ヒットした開始位置までデータを飛ばし、飛ばしたＨＴＭＬ文書をＨＴＭＬ文書一時格納領域（２７１１）に格納する。具体的には、図３４のＨＴＭＬ文書（３４００）で、検索ターム

が「特集記事」とした場合、検索ヒットした「特集」（３４０３）の前に存在する「今月の」までのＨＴＭＬ文書をＨＴＭＬ文書一時格納領域（２７１１）に挿入する。

ステップ４２１０：ステップ４２０５同様に、ハイライト用タグの開始タグをＨＴＭＬ文書一時格納領域（２７１１）に格納する。ここでは、「＜ＢＬＩＮＫ＞」を挿入する。

ステップ４２１１：検索ヒットした文字列を、ＨＴＭＬ文書一時格納領域（２７１１）に挿入する。但し、検索ヒットした文字列の途中にタグが存在する場合は、タグが存在する箇所までの文字列を挿入する。例えば、ＨＴＭＬ文書（３４００）で、検索タームが「特集記事」とした場合、「特集」（３４０３）と「記事」（３４１６）の間に「＜／Ｈ１＞」（３４１７）が存在する。よって、ここでは、「特集」を格納する。

ステップ４２１２：ハイライト用タグの終了タグをＨＴＭＬ文書一時格納領域（２７１１）に挿入する。ここでは、「＜／ＢＬＩＮＫ＞」を挿入する。

【００８０】ステップ４２１３：ステップ４２１１において、検索タームの文字列すべてをＨＴＭＬ文書一時格納領域（２７１１）に挿入したか否かをチェックする。ＨＴＭＬのタグを除くことにより、検索タームとＨＴＭＬ文書の文字列がヒットし、検索ヒットした先頭位置から検索ターム長の文字列の間に、ＨＴＭＬのタグが存在する場合、ステップ４２１１では、ＨＴＭＬタグまでのＨＴＭＬ文書をＨＴＭＬ文書一時格納領域（２７１１）に挿入する。この場合、ＨＴＭＬタグから残りの検索ヒットの文字を処理する必要がある。すべての検索タームをＨＴＭＬ文書一時格納領域に挿入した場合は、処理を終了する。また、ＨＴＭＬタグから残りの検索ヒットの文字を処理する場合は、ステップ４２００に戻る。図３４のＨＴＭＬ文書（３４００）で、検索タームが「特集記事」とした場合、「特集」（３４０３）と「記事」（３４１６）の間に「＜／Ｈ１＞」（３４１７）が存在する。ステップ４２０６では、「特集」のみ挿入した状態で、「記事」を挿入していないため、ステップ４２００に戻る。

【００８１】このような処理を行うことで、クライアント（２７０１）が設定した検索タームを用いて、検索タームと合致するＨＴＭＬ文書にハイライトタグを挿入し、Webブラウザ（２７０３）にハイライトヒット箇所を表示することが可能である。ここでは、検索ターム１つに対して、ＨＴＭＬ文書をチェックし、検索タームの文字列がＨＴＭＬ文書中に存在すれば、クライアント（２７０１）のWebブラウザに検索の結果を表示する処理方法を示したが、1つの検索タームに対して、複数のＨＴＭＬ文書から検索し、検索ヒットしたＨＴＭＬ文書数分のハイライト位置情報を格納し、ハイライト用タグを格納した複数ＨＴＭＬ文書を一括して作成する事も可

能である。また、複数の検索タームに対して、複数のＨＴML 文書から検索し、検索ヒットしたＨＴML 文書数分のハイライト 位置情報を格納し、ハイライト 用タグを格納した複数ＨＴML 文書を一括して作成することも可能である。

【００８２】次に本発明を用いた実施例6 について説明する。本実施例の実施例2 からの変更点は、検索条件中に検索タームなどと共に、検索条件にヒットした場合のハイライト 方法を定義することができる点である。これにより、任意の検索条件に対して、検索条件毎にハイライト 方法を指定することができる。本実施例のシステム構成は図1 と同じである。ただし、検索条件１０３の記載方法が異なる。本実施例における検索条件１０３の記述方法の例を図４３を用いて説明する。

【００８３】図４３に本実施例における検索条件の例を示す。本図に示すように、各検索タームや構造条件などの後ろに、「 {アンダーライン}」のようにハイライト 方法を指定する。実施例2 における検索条件は、「 検索対象の構造指定: 検索条件式」であったが、「 検索対象の構造条件{ハイライト 方法}: ハイライト 方法付き検索条件式」となる。ハイライト 方法の指定は省略可能である。省略時は、実施例2 で示した方法でハイライト 表示を行なう。すなわち、ハイライト 方法が検索条件中に記載されていない箇所については、図１８に示したハイライト 方法定義１８０１を読み出し、本定義情報に記載されているハイライト 方法を用いてハイライト 表示する。

【００８４】図４４に本実施例におけるヒット 範囲情報４４０１の格納内容を示す。実施例2 の図１７に示したヒット 範囲情報からの変更点は、各ヒット 範囲ごとにヒット 条件４４０２だけではなく、ハイライト 方法４４０３を格納する点である。本情報は、図４３を用いて前述した検索条件を解析し、検索条件中に記載されたハイライト 方法の情報を読み出すことで取得可能である。

【００８５】図４５に本実施例におけるハイライト 表示用ＤＴＤの生成方法を示す。本例では、検索の度に新規にハイライト 方法が変更される可能性があることから、ハイライト 表示の度に、必要な構造だけを追加したハイライト 表示用ＤＴＤを生成することとする。この場合、ＤＴＤ中に検索条件ではなく、直接ハイライト 方法に関する記述を行なうことになる。本図に示すように、登録に用いた元のＤＴＤ（１９０１）に対して、上位のハイライト 構造内には下位のハイライト 構造を階層的に指定でき、さらに省略も可能なように定義を変更、追加したハイライト 表示用のＤＴＤ（４５０１）を生成している。

【００８６】ＤＴＤの作成方法は、まず図４４のヒット 範囲情報にハイライト 方法４４０３が記載されていない場合に、図１８に示したハイライト 方法定義からヒット 条件に対応するハイライト 方法を取得する。まず、元のＤＴＤの各構造に対して、下位構造に出現するハイライト 方法を内容モデルに持つことができるように、構造情報を変更する（４５０２）。さらに、ヒット 範囲情報４４０１におけるヒット 範囲の階層関係から、出現するハイライト 用構造の階層関係を得る。ここで得られたハイライト 表示用の階層関係を元に、各ハイライト 構造の下位構造として、下位のハイライト 構造および文字列を内容モデルとして持つようにする。下位のハイライト 構造がなければ、内容モデルとして、文字列だけが出現するようにする（４５０３）。

【００８７】本実施例におけるハイライト 処理により、検索条件をハイライト 構造とするのではなく、記載されたハイライト 方法を記述したハイライト 表示用構造化文書と、ハイライト 表示用のＤＴＤを生成することになる。このように、本実施例による処理により、ハイライト 表示用の構造化文書が生成される。生成されたハイライト 表示用の構造化文書の例を図４６に示す。図４６に示すハイライト 表示用の構造化文書をハイライト 表示すると、図４７に示すようになる。

【００８８】

【 発明の効果】本発明により、構造化文書の検索結果として、ヒットした文書の内容を表示する際に、各構造ごとに検索タームがヒットした範囲に、ハイライト 情報を付加した構造化文書を出力することが可能となる。ブラウザ依存のハイライト 情報ではなく、構造化文書中にハイライト 情報を埋め込むことで、どのようなブラウザにおいてもハイライト 表示が可能となる。検索時の条件、または、各検索タームの重要度、出現頻度などの条件によって異なるハイライト 処理が行え、重要な検索タームについては、高い重み付けであることを明示したハイライト 処理を行なうことが可能となる。さらに、検索条件中にハイライト 方法を記述することで、ユーザ毎に任意のハイライト 表示を行なうことが可能となる。さらに、部分構造だけを抽出して、ハイライト 情報を付加した構造化文書を出力することが可能になる。また、文書構造を示すＨＴMLタグが存在する文書から文字列を検索する場合、設定した検索タームと一致した文字列がＨＴＭＬタグ内に存在する場合や、検索タームがＨＴML タグをまたがって記述されている場合でも容易に検索ヒットすることが可能となる。また、検索ヒットした文字列をハイライト 表示することが可能となる。

【 図面の簡単な説明】

【 図1 】実施例1 、２の構造化文書検索表示装置の処理ブロック図である。

【 図2 】構造化文書検索表示処理のフローチャートを示す図である。

【 図3 】構造化文書登録の内容を示す図である。

【 図4 】構造化文書登録処理のフローチャートを示す図である。

【 図5 】検索用のテキストを示す図である。

【 図6 】更新処理のフローチャートを示す図である。

39

【 図7 】指定構造の抽出処理のフローチャートを示す図である。
【 図8 】構造指定の解析結果として出力される情報を示す図である。
【 図9 】文書表示処理のフローチャートを示す図である。
【 図１０ 】構造化文書およびハイライト処理結果の例を示す図である。
【 図１１ 】文書表示用ＤＴＤ作成処理のフローチャートを示す図である。
【 図１２ 】構造化文書検索用の正規化処理の内容を示す図である。
【 図１３ 】正規化処理を行なった結果の格納内容を示す図である。
【 図１４ 】正規化処理を行なった場合のヒット範囲情報の変換処理内容を示す図である。
【 図１５ 】正規化処理を行なった場合のヒット範囲情報の変換処理のフローチャートを示す図である。
【 図１６ 】ハイライト情報を付加する処理のフローチャートを示す図である。
【 図１７ 】実施例2におけるヒット範囲情報を示す図である。
【 図１８ 】実施例2におけるヒット情報ごとのハイライト方法の定義を示す図である。
【 図１９ 】実施例2のハイライト表示用ＤＴＤへの変換内容を示す図である。
【 図２０ 】実施例2におけるハイライト処理のフローチャートを示す図である。
【 図２１ 】実施例2によりハイライト情報を付加したＳＧＭＬ文書の例を示す図である。
【 図２２ 】ハイライト表示の例を示す図である。
【 図２３ 】実施例3の構造化文書検索表示装置の概略処理ブロック図である。
【 図２４ 】実施例3の処理内容のフローチャートを示す図である。
【 図２５ 】部分構造表示用のＤＴＤへの変換処理を示す図である。
【 図２６ 】部分構造表示用のＤＴＤ作成処理のフローチャートを示す図である。
【 図２７ 】実施例4におけるシステム構成図である。
【 図２８ 】データ制御部のフローチャートを示す図である。
【 図２９ 】実施例4における文字検索処理およびハイライト位置情報の作成処理のフローチャートを示す図である。
【 図３０ 】ハイライト位置情報格納領域の構成である。
【 図３１ 】ハイライト数格納領域の構成である。
【 図３２ 】ハイライトタグ文字格納領域の構成である。
【 図３３ 】実施例4におけるハイライトタグ付きＨＴＭＬ文書の作成処理のフローチャートを示す図である。

40

【 図３４ 】ハイライト挿入例である。
【 図３５ 】ハイライト挿入後の例である。
【 図３６ 】実施例5におけるシステム構成図である。
【 図３７ 】実施例5における検索処理およびハイライト位置情報作成処理2のフローチャートを示す図である。
【 図３８ 】実施例5におけるタグ内検索およびタグ外検索処理のフローチャートを示す図である。
【 図３９ 】実施例5におけるタグ外用検索処理のフローチャートを示す図である。
【 図４０ 】実施例5におけるハイライト用タグの挿入ＨＴＭＬ文書の作成処理のフローチャートを示す図である。
【 図４１ 】実施例5におけるハイライトタグ挿入処理のフローチャートを示す図である。
【 図４２ 】実施例5におけるタグ外ハイライトタグ挿入処理のフローチャートを示す図である。
【 図４３ 】実施例6における検索条件の例である。
【 図４４ 】実施例6におけるヒット範囲情報の例である。
【 図４５ 】実施例6におけるハイライト表示用ＤＴＤへの変換処理を示す図である。
【 図４６ 】実施例6におけるハイライト表示用のＳＧＭＬ文書の例を示す図である。
【 図４７ 】実施例6におけるハイライト表示の例を示す図である。

【 符号の説明】

１０１　構造化文書検索表示装置
１０２　登録用構造化文書
１０３　検索条件
１０４　文書登録処理モジュール
１０５　構造化文書ＤＢ
１０６　検索用情報ＤＢ
１０７　構造化文書読み出し処理モジュール
１０８　検索処理モジュール
１０９　ヒット文書番号情報
１１０　ヒット範囲情報
１１１　ヒット文書の文書内容
１１２　文書表示処理モジュール
１１３　表示用文書
１１４　登録用文書格納ファイル
１１５　入出力装置
２３０１　表示構造情報
２３０２　部分構造表示モジュール
２７００　ＷＷＷ検索システム
２７０１　クライアント
２７０３　Webブラウザ
２７０４　ＨＴＴＰサーバ
２７０５　データ制御
２７０６　メモリ
２７０７　ＨＴＭＬ文書

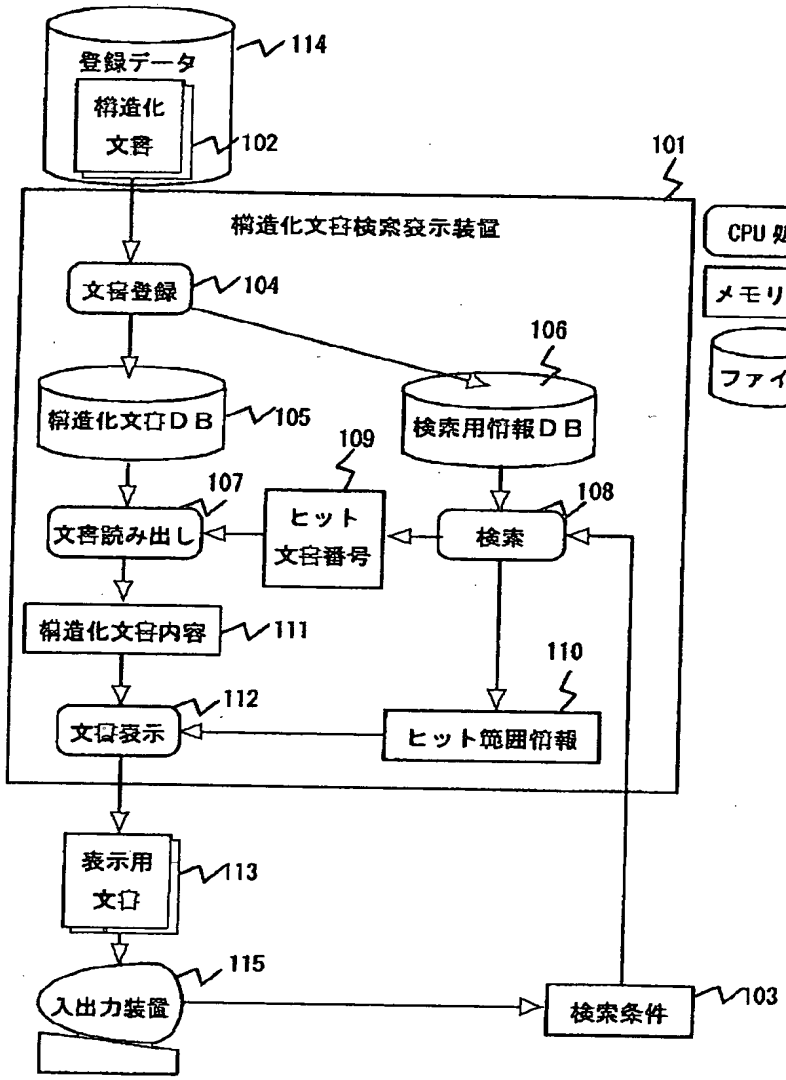
41

３２００　ハイライトタグ文字格納領域の構造
３４００　ＨＴＭＬ文書例
３４０１　ＨＴＭＬ文書の表示画面例

42

３５００　ハイライトタグ挿入後ＨＴＭＬ文書例
３５０１　ハイライトタグ挿入後表示画面例

【図１】



<実施例１、２の処理ブロック図>

【図５】

【図５】

<検索用のテキスト>

構造ＩＤ	開始文字位置	文字列長さ
2	0	10
5	10	6
6	16	44
:	:	:

<文字列>
構造化文書構造化文書とは、・・・・

【図８】

【図８】

<構造指定の解析結果>

構造数	N

601

N

構造ＩＤ
2
3
4
7
8
:

602

【図１４】

【図１４】

<ヒット範囲情報>

正規化後構造ＩＤ	ヒット範囲	
	開始位置	長さ
2	0	10
9	0	10

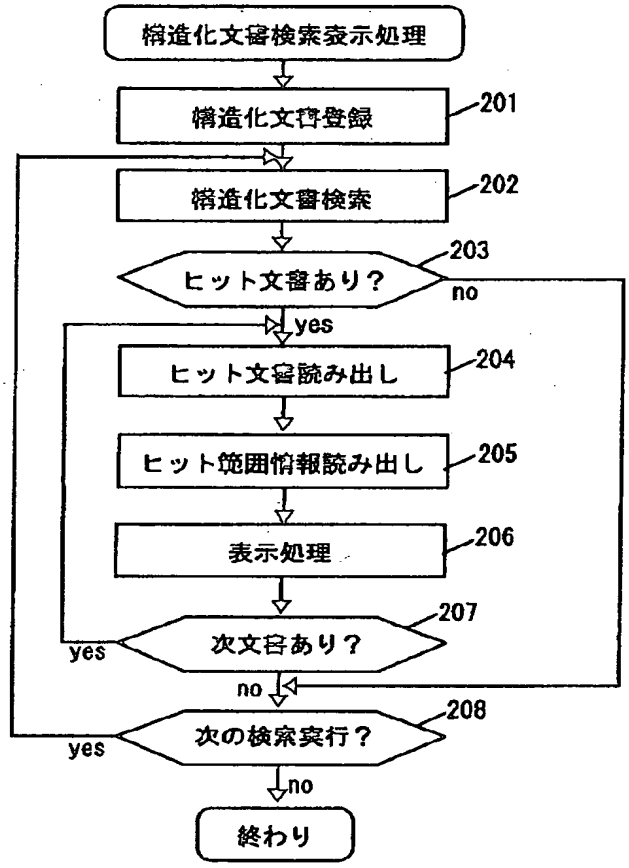
1401

正規化前の範囲への変換

正規化前構造ＩＤ	ヒット範囲	
	開始位置	長さ
2	0	10
5	0	6
6	0	4

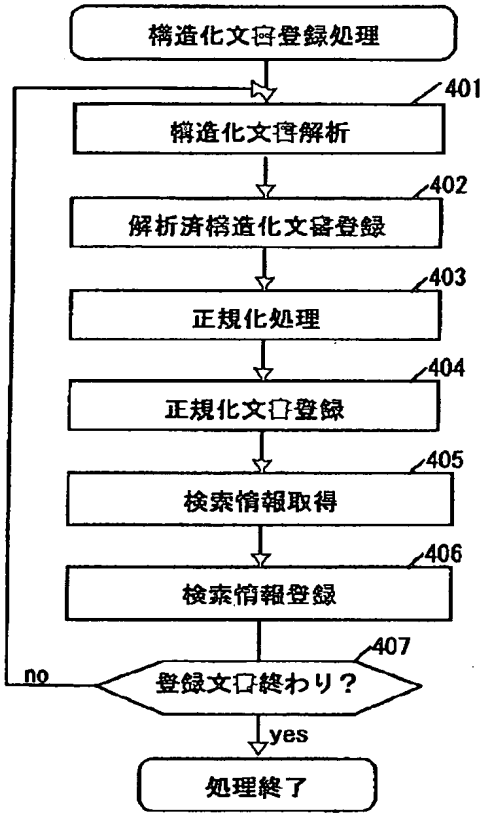
1402

【図２】



＜実施例１の文書表示処理のフローチャート＞

【図４】



＜構造化文書登録処理のフローチャート＞

【図１７】

【図17】

＜ヒット範囲情報＞

構造ＩＤ	ヒット条件	ヒット範囲	
		開始位置	長さ
2	ターム	0	10
9	頻度	0	10
9	構造	0	50
9	ターム	30	4
9	ターム	36	4
9	距離	30	10

1701
1702

【図１８】

【図18】

＜ハイライト方法定義＞

ヒット条件	ハイライト方法	範囲情報
ターム	赤色	1
頻度	フォント大	1
距離	アンダーライン	2
構造	斜体	3

1801
1802
1803
1804

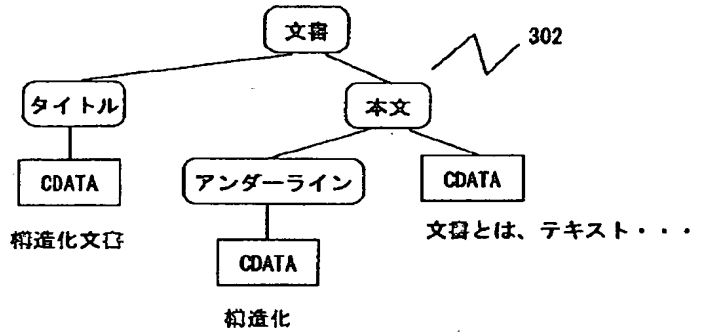
【図３】

【図３】

<構造化文書の登録>

301

```
<!DOCTYPE 文書 SYSTEM "文書.dtd">
<文書>
<タイトル>構造化文書</タイトル>
<本文>
<アンダーライン>構造化</アンダーライン>文書とは、テキスト・・・
</本文>
</文書>
```



テーブル形式の
データへの変換

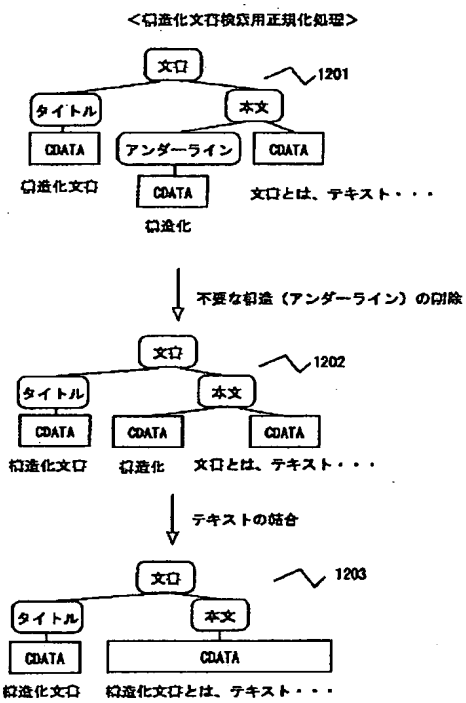
303

構造識別子	構造種別	タグ	内容
0	文書	文書	下位構造：1,3
1	要素	タイトル	下位構造：2
2	CDATA	—	文字列:"構造化文書"
3	要素	本文	下位構造：4,6
4	要素	アンダーライン	下位要素：5
5	CDATA	—	文字列:"構造化"
6	CDATA	—	文字列:"文書とは、・・"

【図１２】

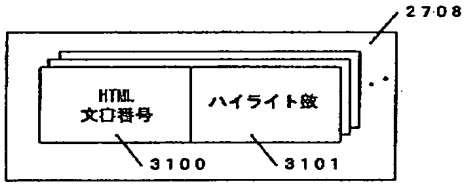
【図12】

<構造化文書検索用正規化処理>



【図３１】

【図31】



【図２２】

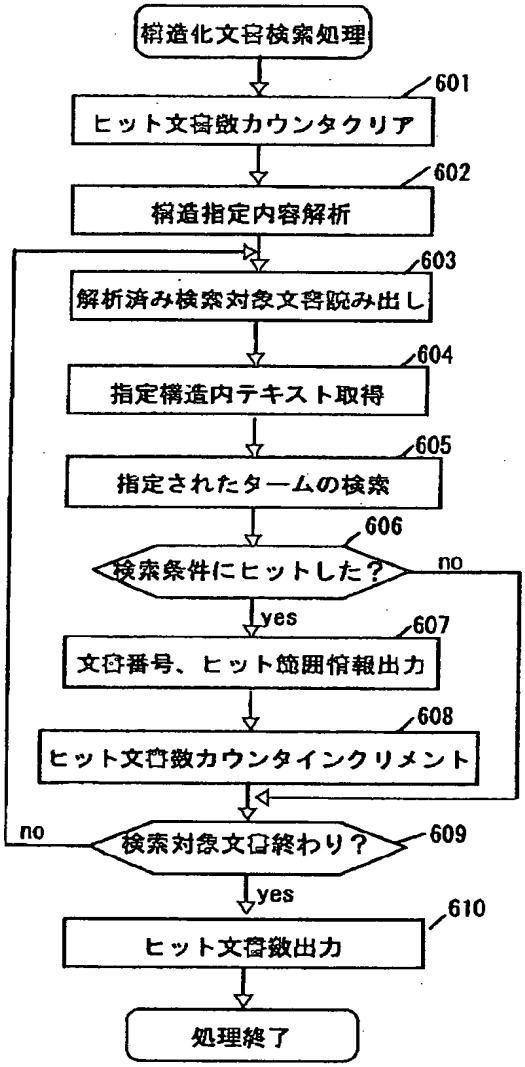
【図２２】

<ハイライト表示の例>

構造化文書とは、テキスト中にタグを挿入した文書・・・・

【 図6 】

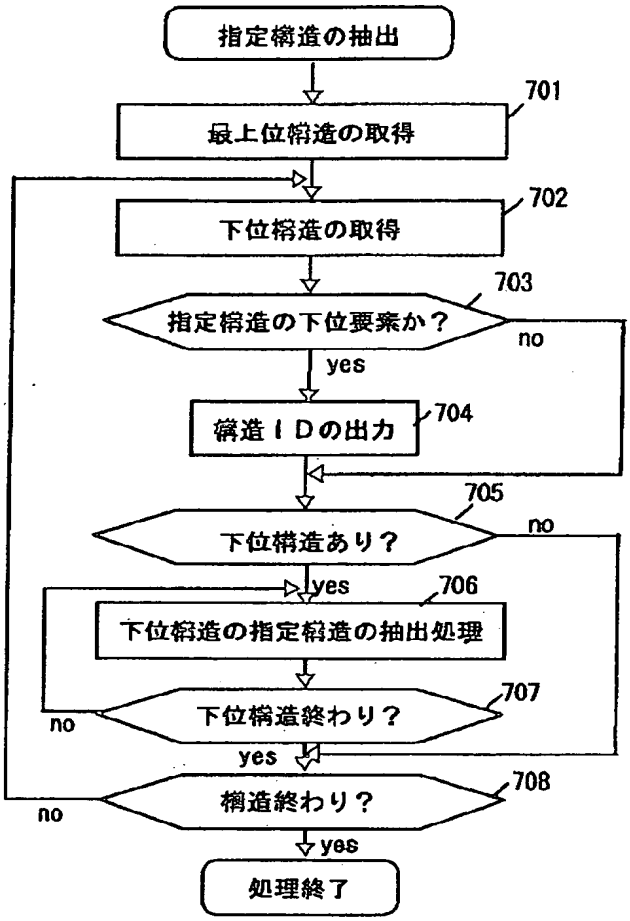
【図６】



<構造化文書検索処理のフローチャート>

【 図7 】

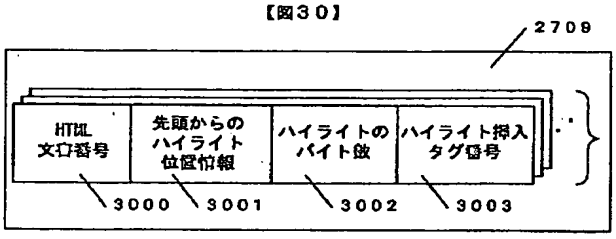
【図７】



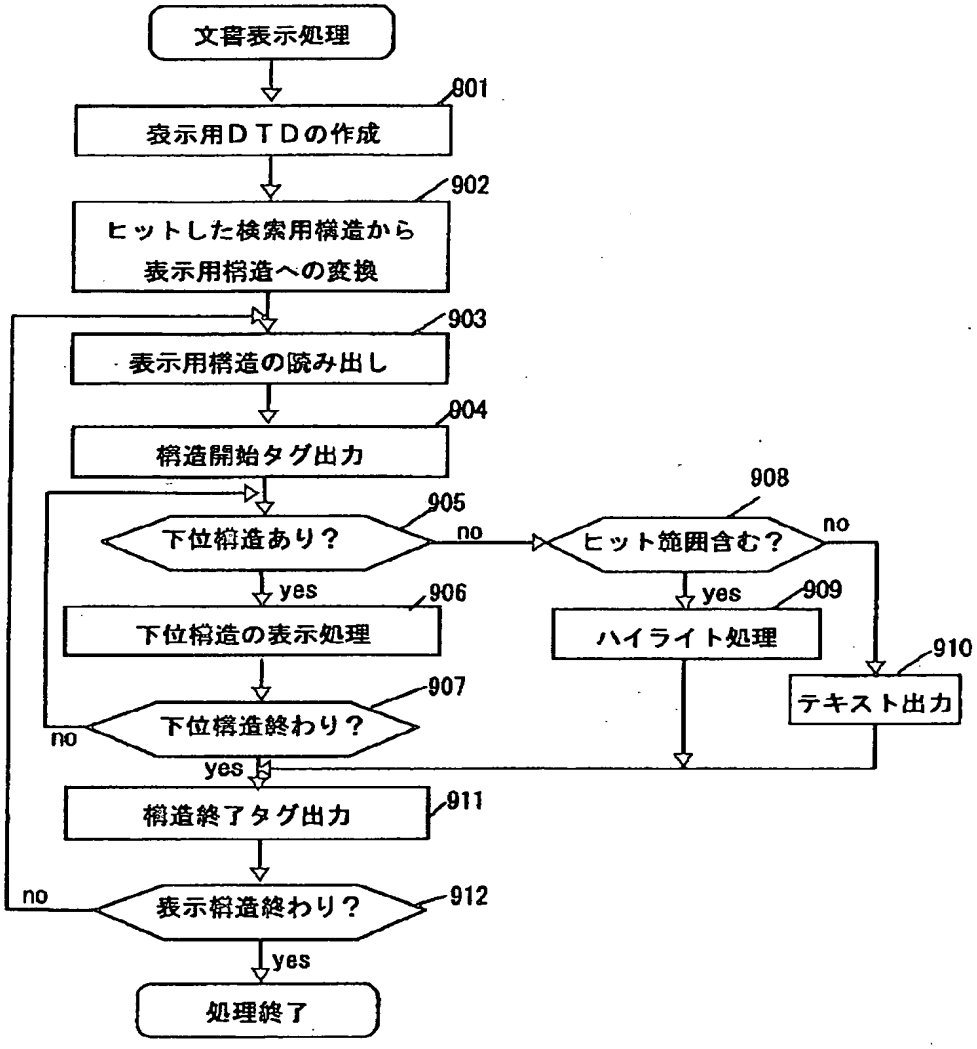
<指定構造抽出処理のフローチャート>

【 図３０ 】

【図３０】



【図9】



＜構造化文書表示処理のフローチャート＞

【図４３】

【図４３】

```
<文書.タイトル>:"構造化文書"[背色] AND
<文書.本文>[波下線]:("構造化文書" OR C<10[反転]("タグ","挿入"))
```

【図１０】

【図１０】

＜構造化文書の例＞

[ＤＴＤ] (文書.dtd)　　1001

```
<!ELEMENT 文書          - - (タイトル,本文)>
<!ELEMENT タイトル      - - CDATA>
<!ELEMENT 本文          - - (#PCDATA|アンダーライン)*>
<!ELEMENT アンダーライン - - CDATA>
```

[文書インスタンス]　　1002

```
<!DOCTYPE 文書 SYSTEM "文書.dtd">
<文書>
<タイトル>構造化文書</タイトル>
<本文>
<アンダーライン>構造化</アンダーライン>文書とは、テキスト、、、
</本文>
</文書>
```

↓ "構造化文書"を検索、結果表示

＜ハイライト処理後の構造化文書の例＞

[ＤＴＤ](表示.dtd)　　1003

```
<!ELEMENT 文書          - - (タイトル,本文)>
<!ELEMENT タイトル      - - (#PCDATA|ハイライト)*>
<!ELEMENT 本文          - - (#PCDATA|アンダーライン|ハイライト)*>
<!ELEMENT アンダーライン - - (#PCDATA|ハイライト)*>
<!ELEMENT ハイライト    - - (#PCDATA)>
```

[文書インスタンス]　　1004

```
<!DOCTYPE 文書 SYSTEM "表示.dtd">
<文書>
<タイトル><ハイライト>構造化文書</ハイライト></タイトル>
<本文>
<アンダーライン><ハイライト>構造化</ハイライト></アンダーライン>
<ハイライト>文書</ハイライト>とは、テキスト、、、
</本文>
</文書>
```

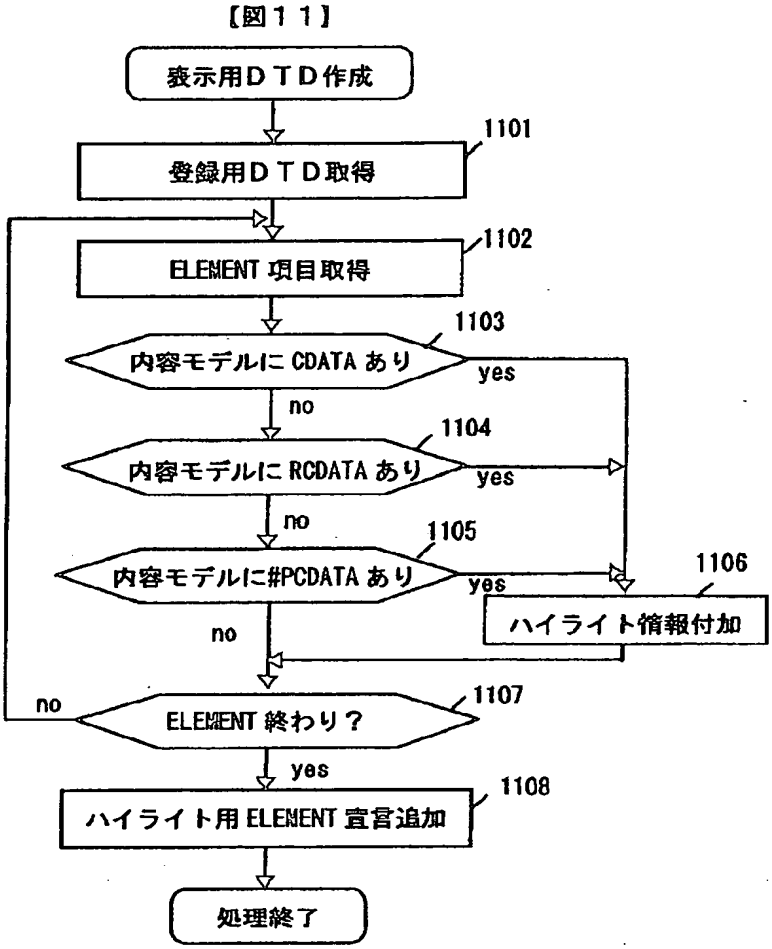
【図４７】

【図４７】

＜ハイライト表示の例＞

構造化文書とは、テキスト中に[文書]を[構造]した文書・・・・

【図１１】



＜表示用ＤＴＤ作成のフローチャート＞

【図４４】

【図４４】

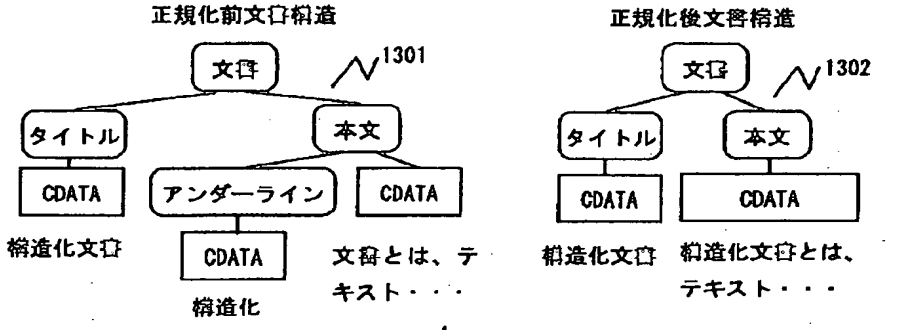
＜ヒット箇所情報＞

構造ＩＤ	ハイライト方法 (4403)	ヒット条件 (4402)	ヒット箇所 (4401)	
			開始位置	長さ
2	背色	ターム	0	10
9	－	頻度	0	10
9	終下位	構造	0	50
9	－	ターム	30	4
9	－	ターム	36	4
9	反転	距離	30	10

【図１３】

【図１３】

＜正規化文書のテーブル形式のデータへの変換＞



テーブル形式のデータへの変換

構造情報テーブル　1303

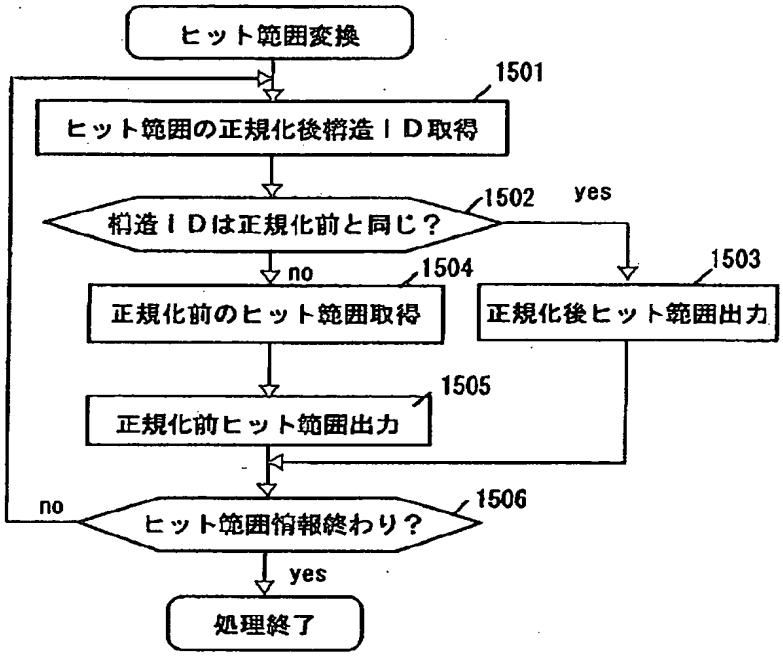
	構造識別子	構造種別	タグ	内容
	0	文書	文書	下位構造：1,3
	1	要素	タイトル	下位構造：2
	2	CDATA	—	文字列:"構造化文書"
	3	要素	本文	下位構造：4,6
	4	要素	アンダーライン	下位構造：5
	5	CDATA	—	文字列:"構造化"
	6	CDATA	—	文字列:"文書とは、・・"
正規化後追加	7	文書	文書	下位構造:1,8
正規化後追加	8	要素	本文	下位構造:9
正規化後追加	9	CDATA	—	文字列:"構造化文書とは、・・"

正規化対応テーブル　1304

正規化後構造識別子	構造種別	正規化前構造識別子	文字列範囲	
			正規化後	正規化前
7	文書	0	—	
8	要素	3	—	
9	CDATA	5	0-5	0-5
		6	6-49	0-43

【図１５】

【図１５】



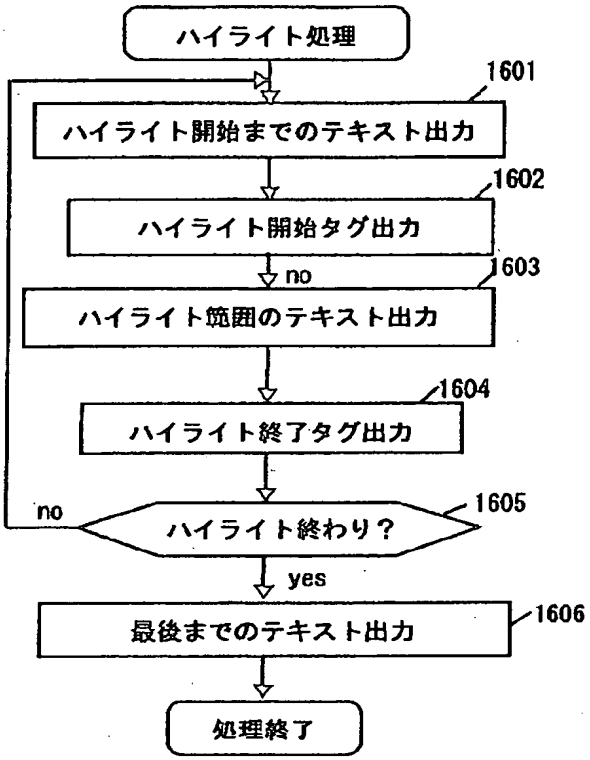
＜ヒット範囲取得処理のフローチャート＞

【図２１】

【図２１】

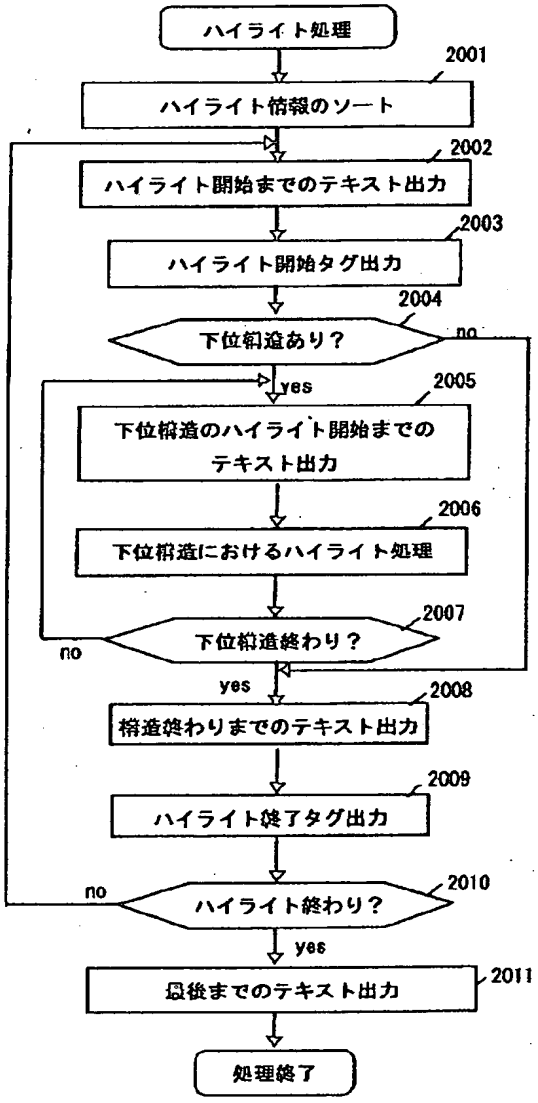
```
<文⊡>
<タイトル><ターム>構造化文⊡</ターム></タイトル>
<本文>
<アンダーライン><構造><頻度>構造化</頻度></構造></アンダーライン><構造><頻度>文⊡</頻度>とは、テキスト中に<距離><ターム>タグ</ターム>を<ターム>挿入</ターム></距離>した文⊡・・・・</構造>
</本文>
</文⊡>
```

【図１６】



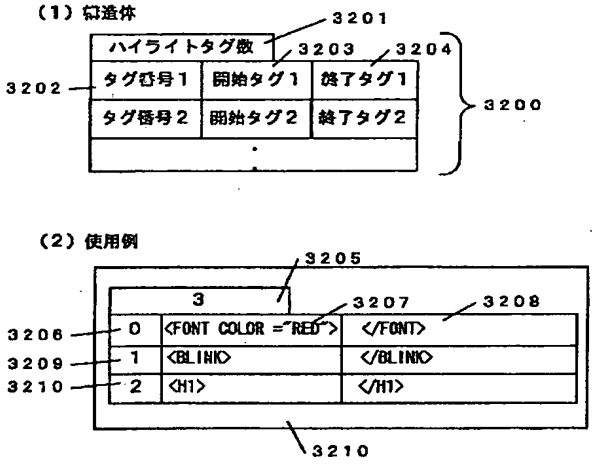
<ハイライト処理のフローチャート>

【図２０】



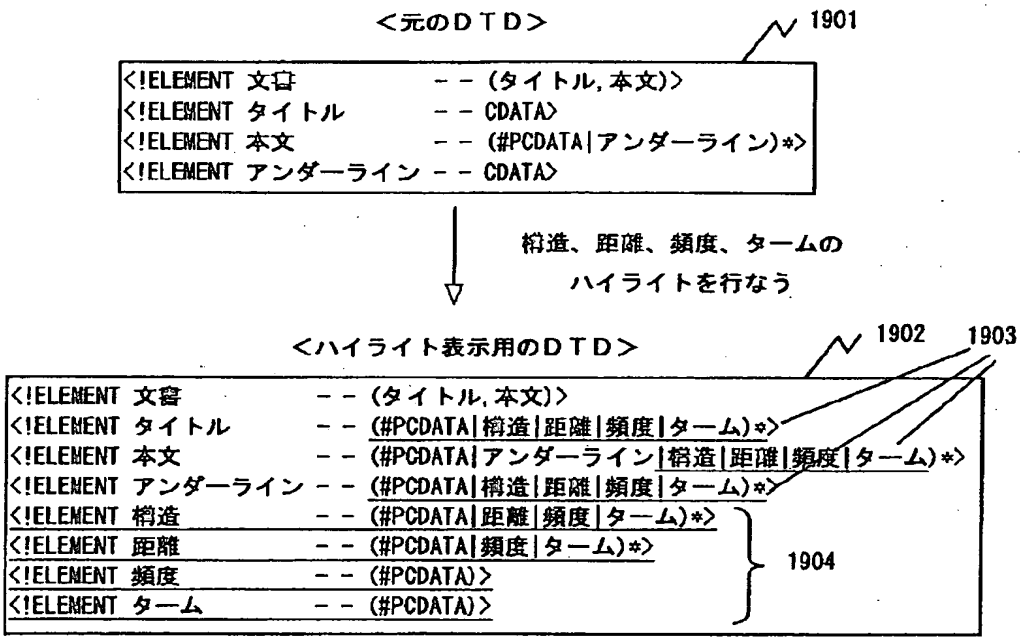
<ヒット範囲取得処理のフローチャート>

【図３２】



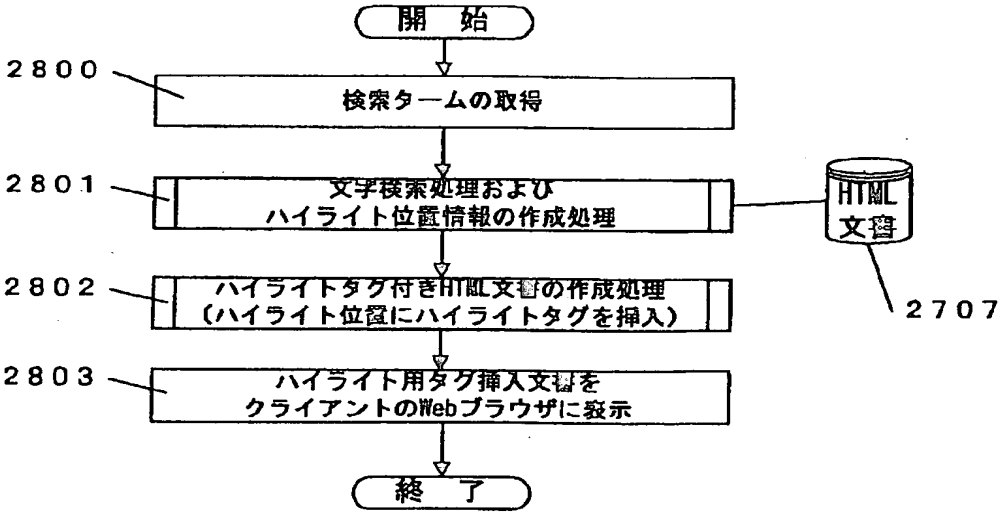
【 図１９ 】

【図１９】



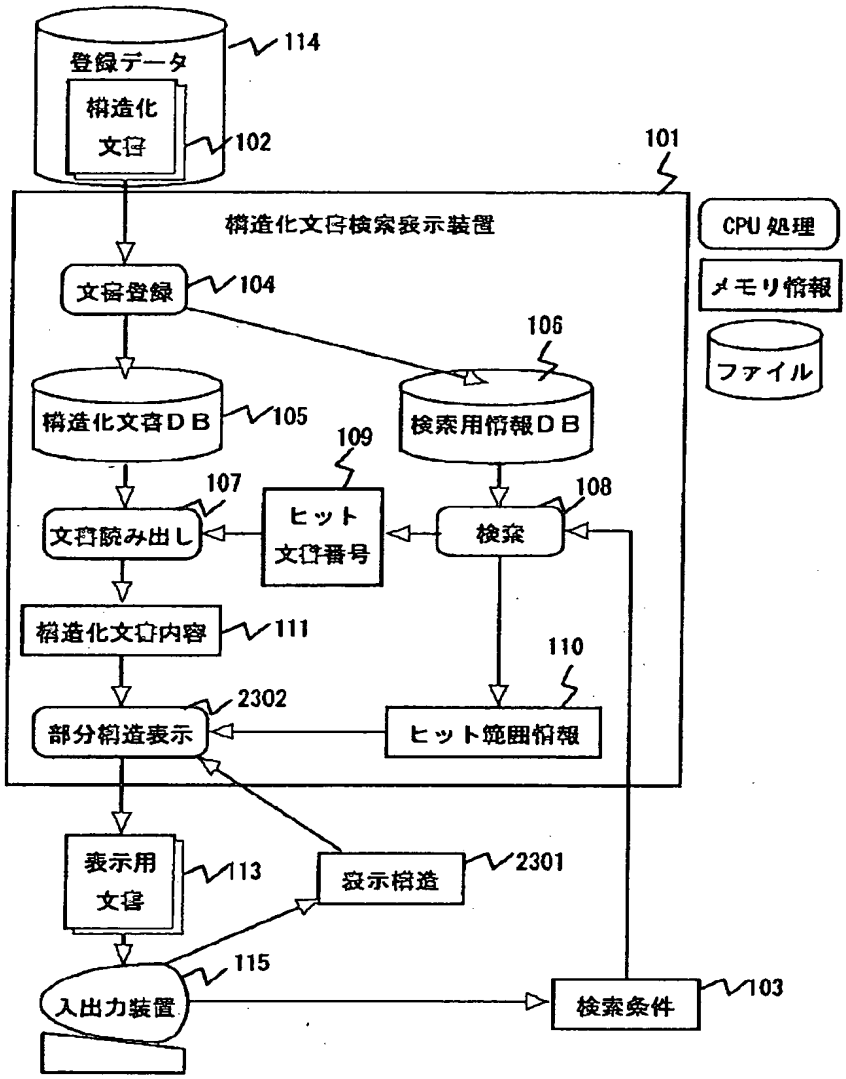
【 図２８ 】

【図２８】



【 図２３ 】

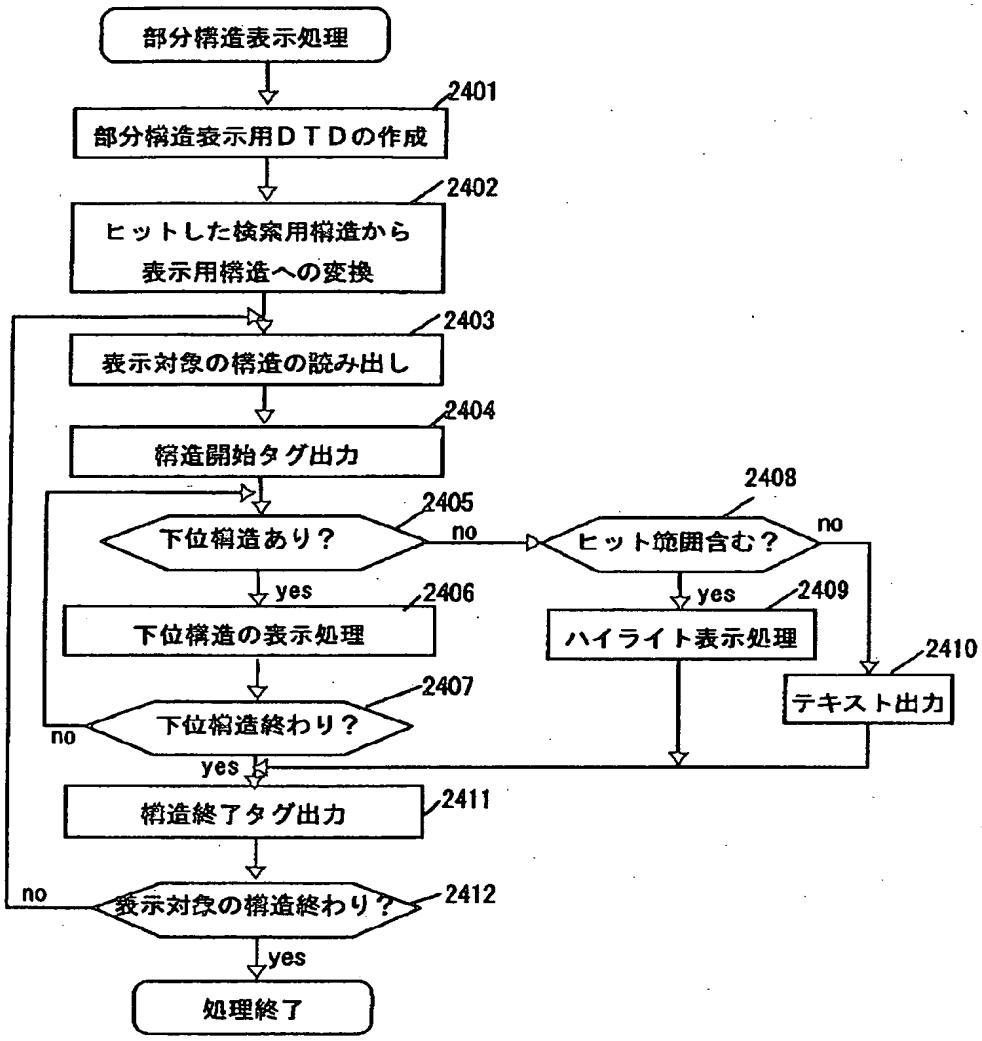
【図２３】



＜実施例３の処理ブロック図＞

【図２４】

【図２４】



＜構造化文書表示処理のフローチャート＞

【図２５】

【図２５】

＜部分構造表示用ＤＴＤへの変換＞

＜元のＤＴＤ＞　2501

```
<!ELEMENT 文書          - - (タイトル, 本文)>
<!ELEMENT タイトル      - - CDATA>
<!ELEMENT 本文          - - (#PCDATA|アンダーライン)*>
<!ELEMENT アンダーライン - - CDATA>
```

部分構造の抽出を可能とする。
ハイライト付加

＜部分構造表示用のＤＴＤ＞　2502

```
<!ELEMENT 文書          O O (タイトル?, 本文?)>
<!ELEMENT タイトル      - - (#PCDATA|ハイライト)*>
<!ELEMENT 本文          O O (#PCDATA|アンダーライン|ハイライト)*>
<!ELEMENT アンダーライン - - (#PCDATA|ハイライト)*>
<!ELEMENT ハイライト    - - (#PCDATA)>
```

＜部分構造を抽出したＳＧＭＬの例＞　2503

```
<タイトル>
<ハイライト>構造化文書</ハイライト>
</タイトル>
```

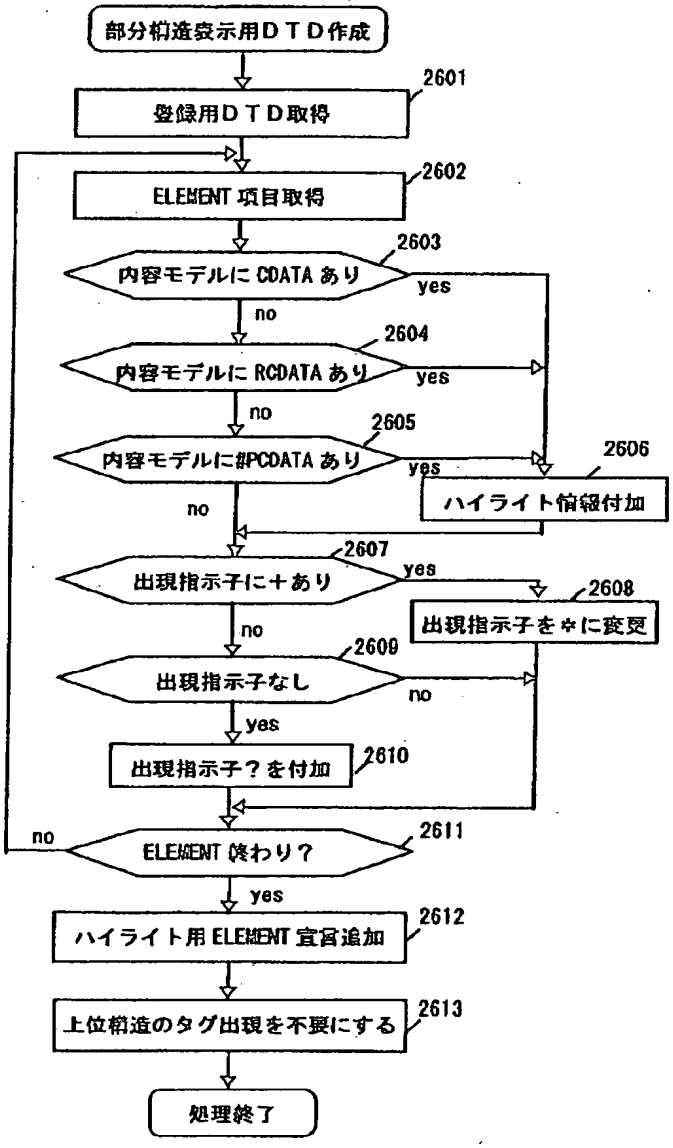
【図４６】

【図４６】

```
<文書>
<タイトル><赤色>構造化文書</赤色></タイトル>
<本文>
<アンダーライン><波下線><フォント大>構造化</フォント大></波下線></アンダーライン><波下線><フォント大>文書</フォント大>とは、テキスト中に<反転><赤色>タグ</赤色>を<赤色>挿入</赤色></反転>した文書・・・・</波下線>
</本文>
</文書>
```

【 図２６ 】

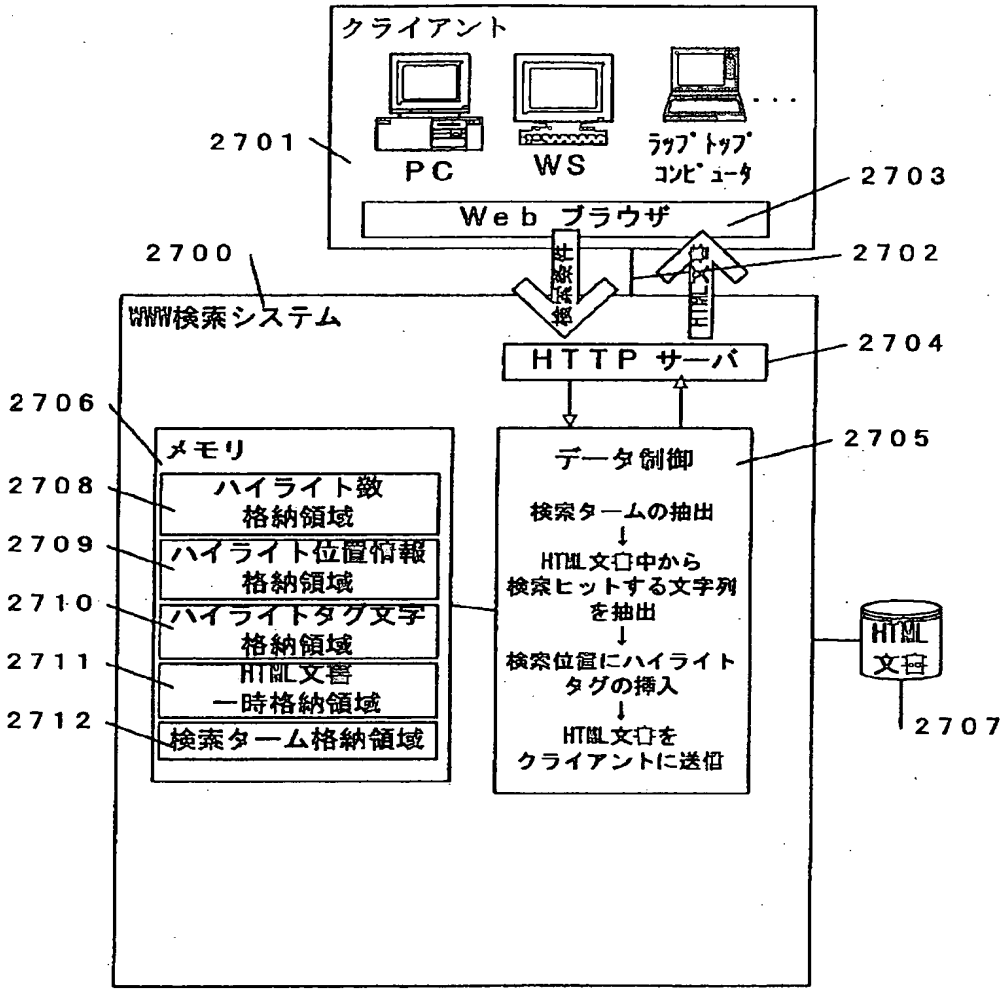
【図２６】



＜部分構造表示用ＤＴＤ作成のフローチャート＞

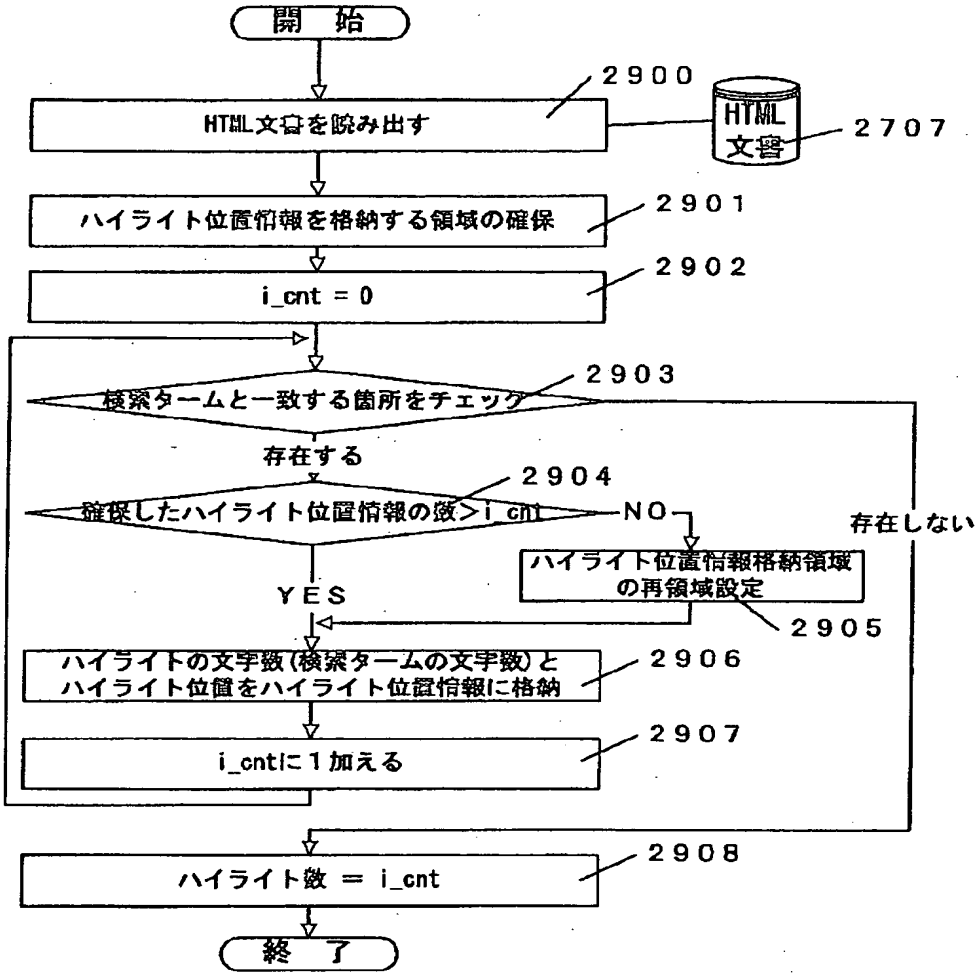
【 図２７ 】

【図２７】


クライアント
PC
WS
ラップトップ コンピュータ
Web ブラウザ
2701
2703
2700
2702
WWW検索システム
検索要求
HTML文書
HTTP サーバ
2704
2706
メモリ
2705
データ制御
2708
ハイライト数 格納領域
2709
ハイライト位置情報 格納領域
2710
ハイライトタグ文字 格納領域
2711
HTML文書 一時格納領域
2712
検索ターム格納領域
検索タームの抽出
↓
HTML文書中から検索ヒットする文字列を抽出
↓
検索位置にハイライトタグの挿入
↓
HTML文書をクライアントに送信
HTML 文書
2707

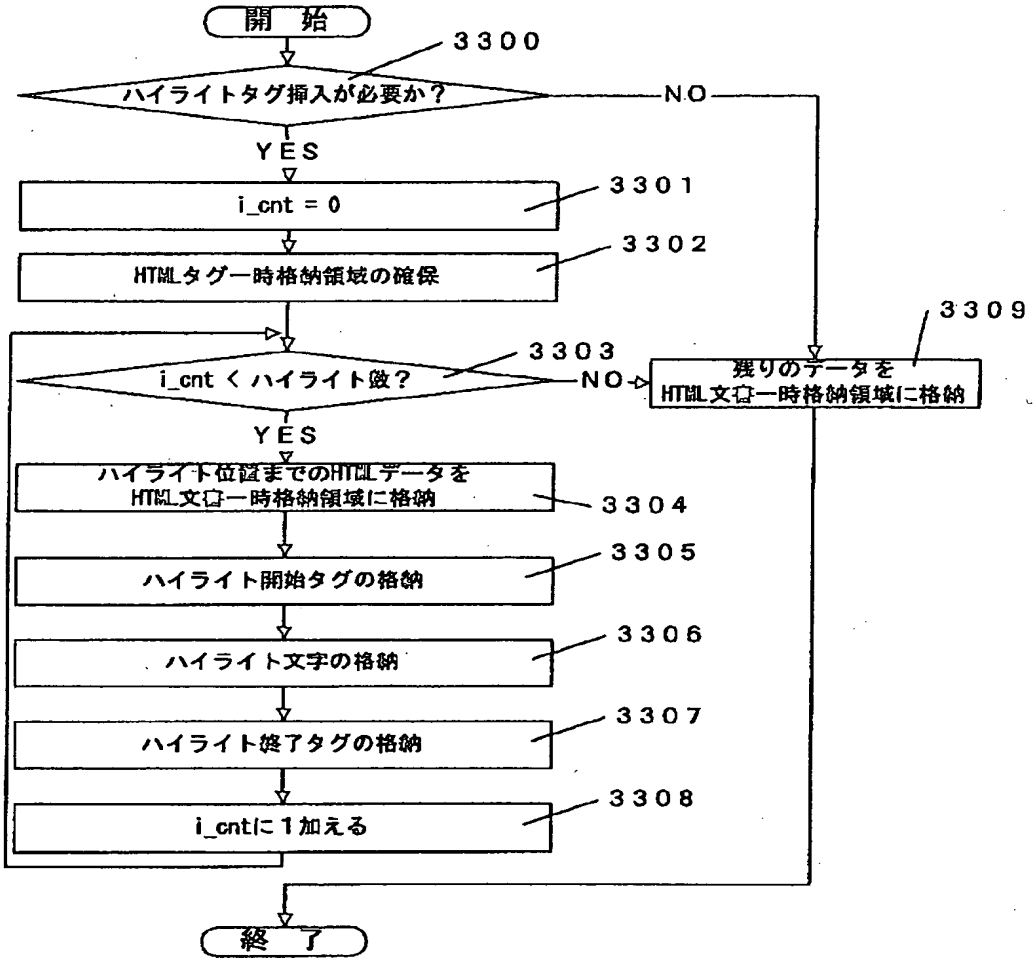
【 図2９ 】

【図２９】


開　始
2900
HTML文書を読み出す
HTML文書
2707
ハイライト位置情報を格納する領域の確保
2901
i_cnt = 0
2902
検索タームと一致する箇所をチェック
2903
存在する
確保したハイライト位置情報の数＞i_cnt
2904
NO
存在しない
ハイライト位置情報格納領域の再領域設定
2905
YES
ハイライトの文字数(検索タームの文字数)とハイライト位置をハイライト位置情報に格納
2906
i_cntに１加える
2907
ハイライト数 ＝ i_cnt
2908
終　了

【 図３３ 】

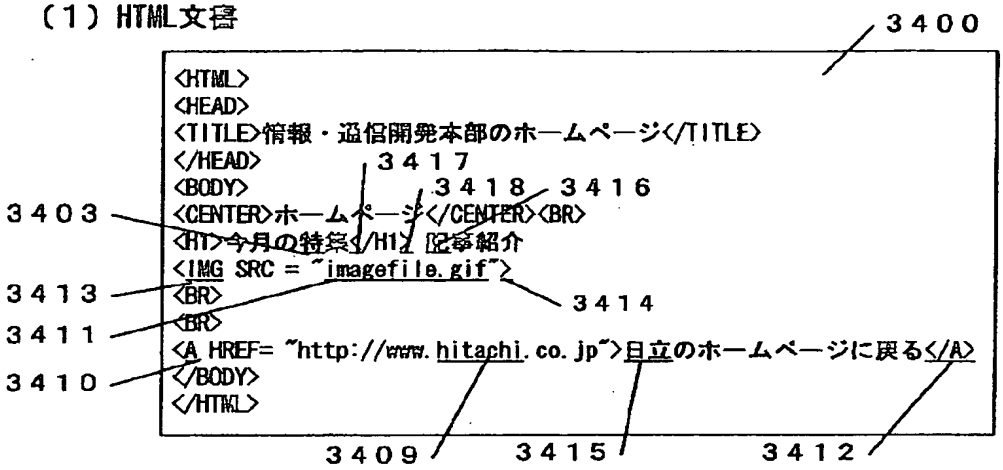
【図３３】


開　始
3300
ハイライトタグ挿入が必要か？
NO
YES
i_cnt = 0
3301
HTML タグ一時格納領域の確保
3302
3309
i_cnt < ハイライト数？
3303
NO
残りのデータを
HTML 文字一時格納領域に格納
YES
ハイライト位置までのHTML データを
HTML 文字一時格納領域に格納
3304
ハイライト開始タグの格納
3305
ハイライト文字の格納
3306
ハイライト終了タグの格納
3307
i_cntに１加える
3308
終　了

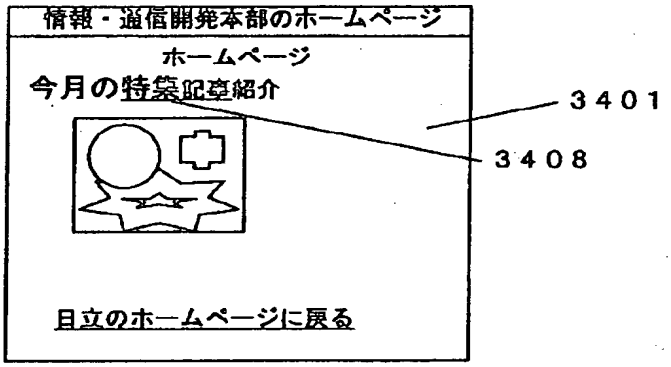
【図３４】

【図３４】

（１）HTML文書



（２）表示画面



（３）ハイライト位置情報格納領域

HTML文書番号	先頭からのハイライト位置情報	ハイライトのバイト数	ハイライト挿入タグ番号
００１	１２２	４	１

３４０２

３４０４　３４０５　３４０６　３４０７

【図３５】

【図３５】

（１）ハイライトタグ挿入後HTML文書

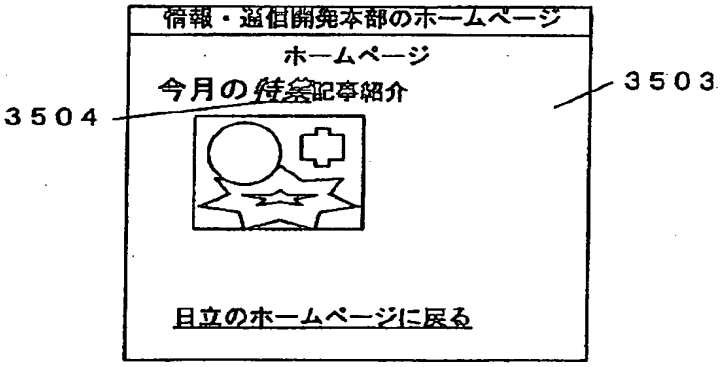
3500

```
<HTML>
<HEAD>
<TITLE>情報・通信開発本部のホームページ</TITLE>
</HEAD>
<BODY>
<CENTER>ホームページ</CENTER><BR>
<H1>今月の<BLINK>特集</BLINK></H1>記事紹介
<IMG SRC = "imagefile.gif">
<BR>
<BR>
<A HREF= "http://www.hitachi.co.jp">日立のホームページに戻る
</A>
</BODY>
</HTML>
```

3501

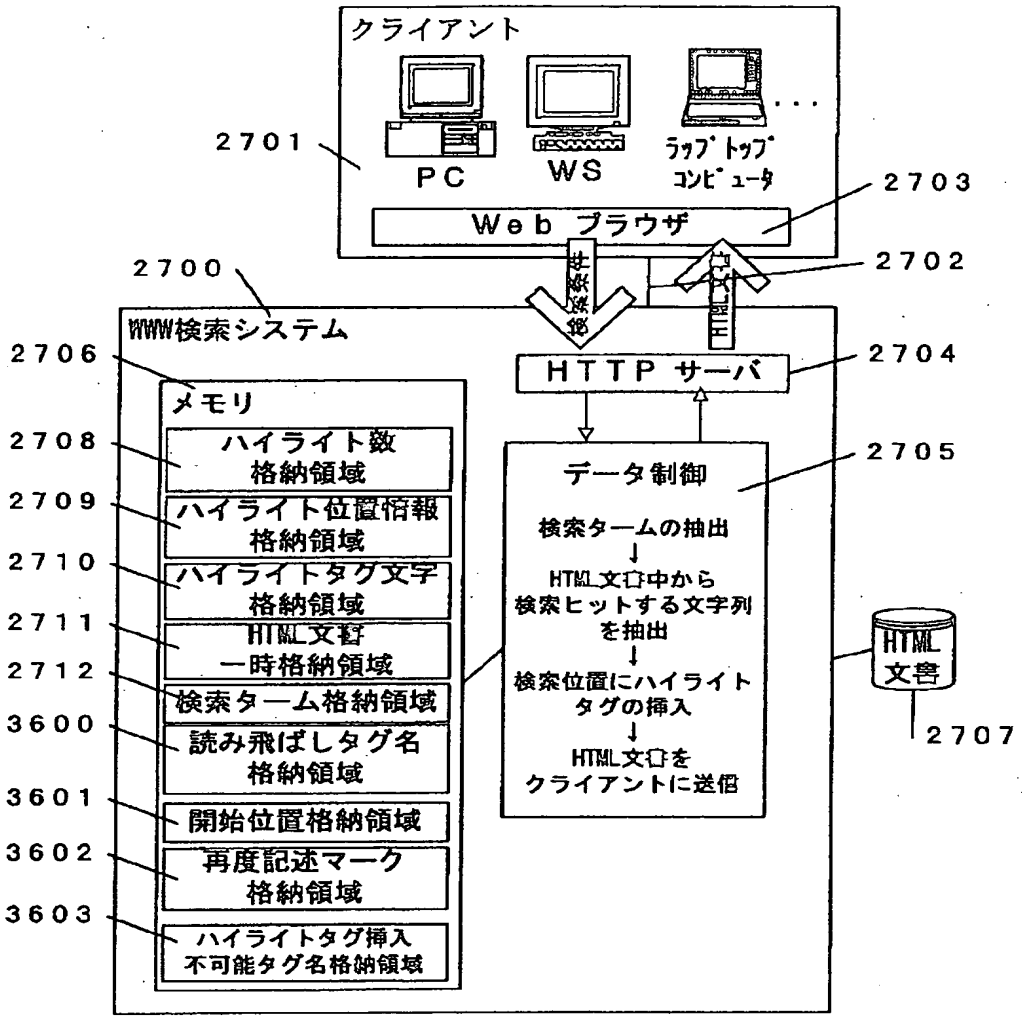
3502

（２）ハイライトタグ挿入後表示画面



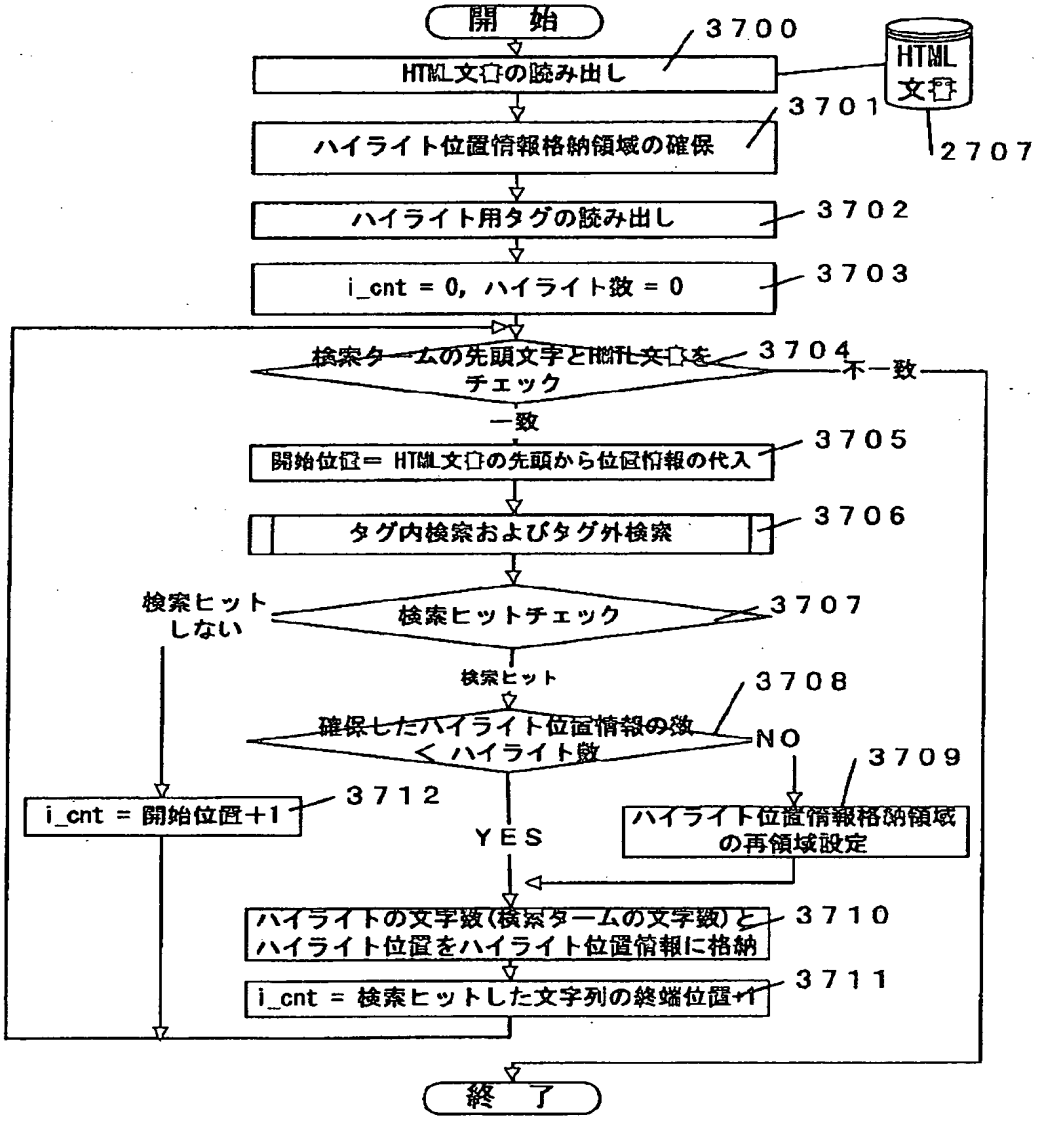
【 図３６ 】

【図３６】


クライアント
2701
PC
WS
ラップトップ コンピュータ
Web ブラウザ
2703
2700
2702
HTTP
WWW検索システム
2706
メモリ
HTTP サーバ
2704
2708
ハイライト数 格納領域
2709
ハイライト位置情報 格納領域
2710
ハイライトタグ文字 格納領域
2711
HTML文書 一時格納領域
2712
検索ターム格納領域
3600
読み飛ばしタグ名 格納領域
3601
開始位置格納領域
3602
再度記述マーク 格納領域
3603
ハイライトタグ挿入 不可能タグ名格納領域
データ制御
2705
検索タームの抽出
↓
HTML文書中から 検索ヒットする文字列 を抽出
↓
検索位置にハイライト タグの挿入
↓
HTML文書を クライアントに送信
HTML 文書
2707

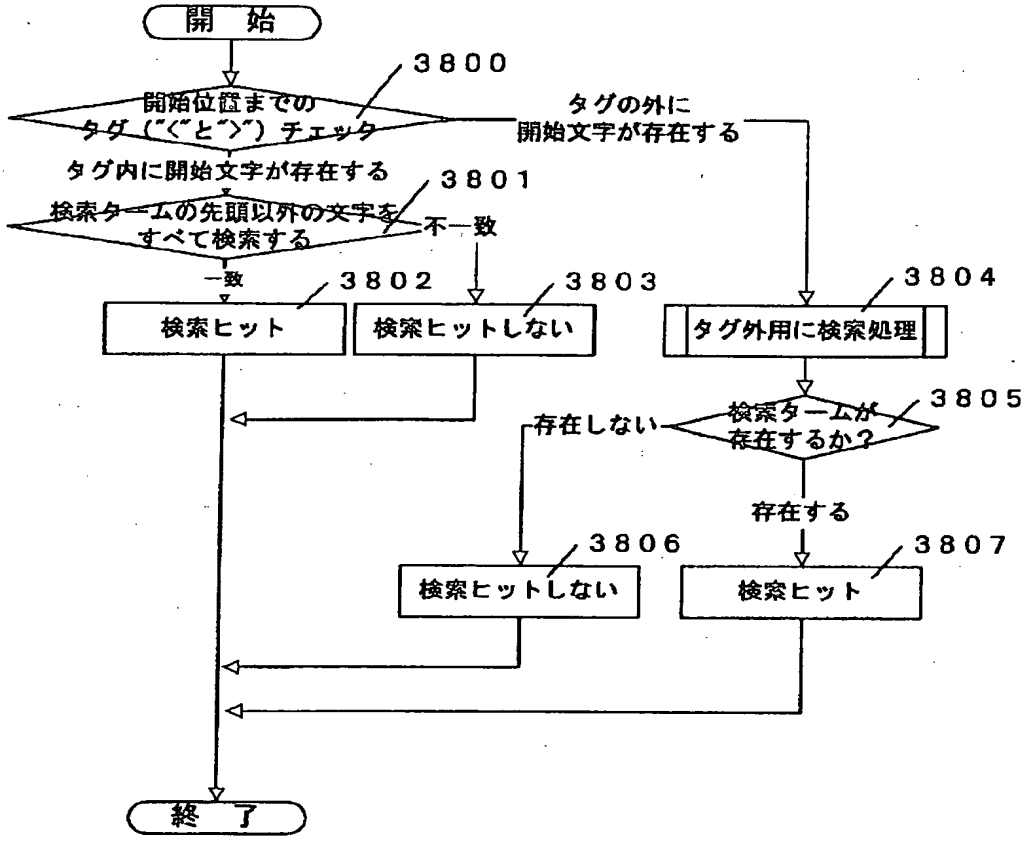
【 図３７ 】

【図３７】

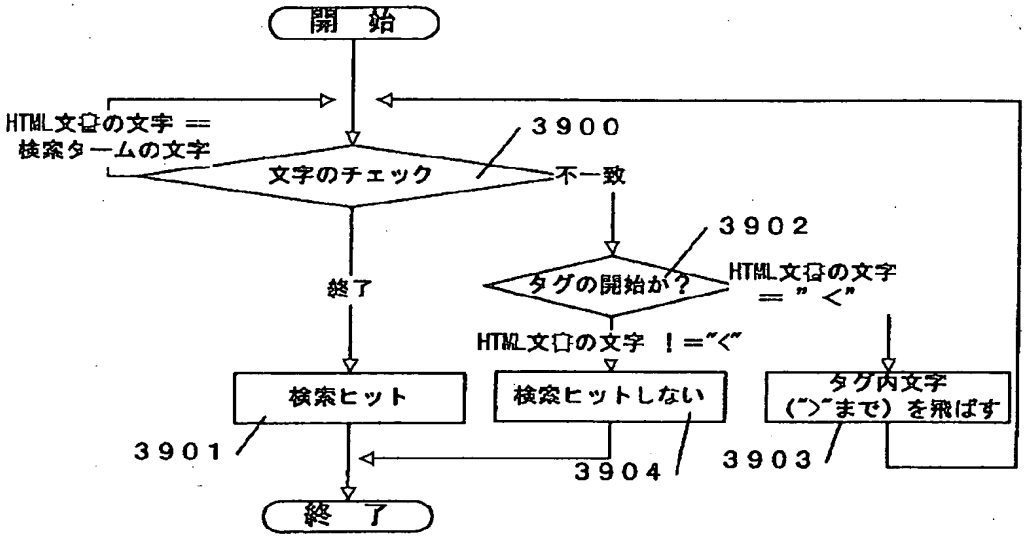

開　始
3700
HTML文書の読み出し
HTML文書
2707
3701
ハイライト位置情報格納領域の確保
3702
ハイライト用タグの読み出し
3703
i_cnt = 0, ハイライト数 = 0
3704
検索タームの先頭文字とHTML文書をチェック
不一致
一致
3705
開始位置＝ HTML文書の先頭から位置情報の代入
3706
タグ内検索およびタグ外検索
検索ヒットしない
検索ヒットチェック
3707
検索ヒット
3708
確保したハイライト位置情報の数 < ハイライト数
NO
3709
ハイライト位置情報格納領域の再領域設定
3712
i_cnt = 開始位置＋1
YES
ハイライトの文字数(検索タームの文字数)とハイライト位置をハイライト位置情報に格納
3710
i_cnt = 検索ヒットした文字列の終端位置＋1
3711
終　了

【 図38 】

【図38】


開　始
3800
開始位置までの
タグ（"<"と">"）チェック
タグの外に
開始文字が存在する
タグ内に開始文字が存在する
3801
検索タームの先頭以外の文字を
すべて検索する
不一致
一致
3802
検索ヒット
3803
検索ヒットしない
3804
タグ外用に検索処理
3805
検索タームが
存在するか？
存在しない
存在する
3806
検索ヒットしない
3807
検索ヒット
終　了

【図３９】



【図４５】

＜実施例６におけるハイライト表示用ＤＴＤへの変換＞

＜元のＤＴＤ＞

1901

```
<!ELEMENT 文書          - - (タイトル,本文)>
<!ELEMENT タイトル      - - CDATA>
<!ELEMENT 本文          - - (#PCDATA|アンダーライン)*>
<!ELEMENT アンダーライン - - CDATA>
```

表示文書用のハイライト情報の付加

＜表示用文書のＤＴＤ＞

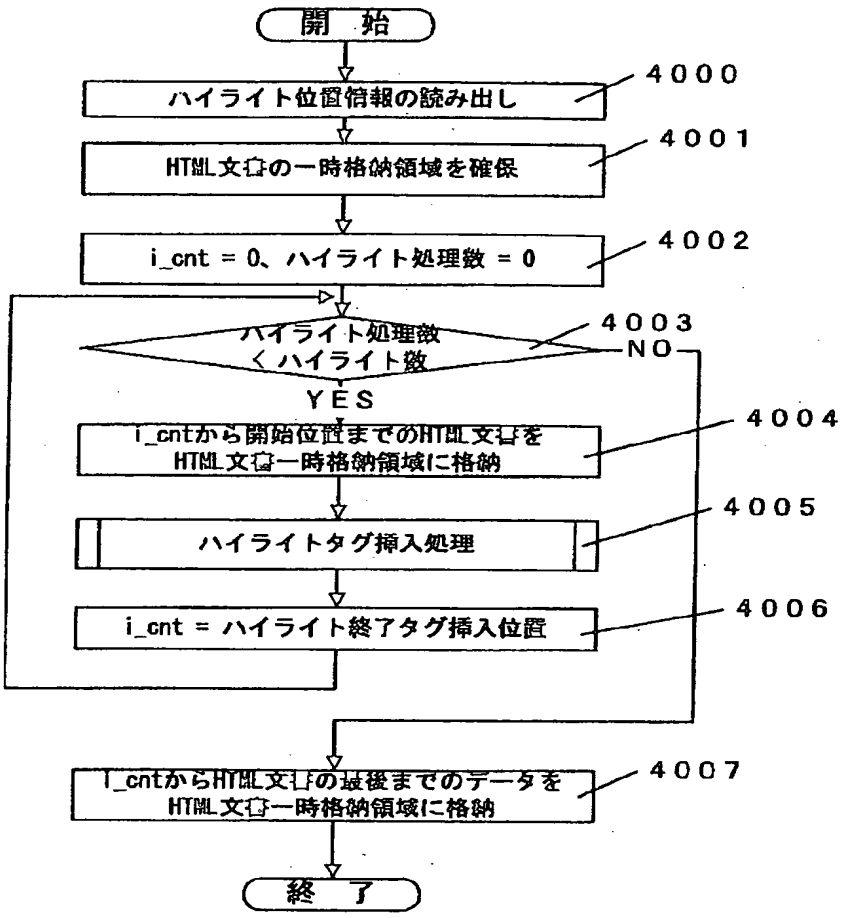
4501　4502

```
<!ELEMENT 文書          - - (タイトル,本文)>
<!ELEMENT タイトル      - - (#PCDATA|背色)*>
<!ELEMENT 本文          - - (#PCDATA|アンダーライン|波下線|フォント大|反転)*>
<!ELEMENT アンダーライン - - (#PCDATA|フォント大)*>
<!ELEMENT 波下線        - - (#PCDATA|フォント大|反転)*>
<!ELEMENT 反転          - - (#PCDATA|赤色)*>
<!ELEMENT フォント大    - - (#PCDATA)>
<!ELEMENT 背色          - - (#PCDATA)>
<!ELEMENT 赤色          - - (#PCDATA)>
```

4503

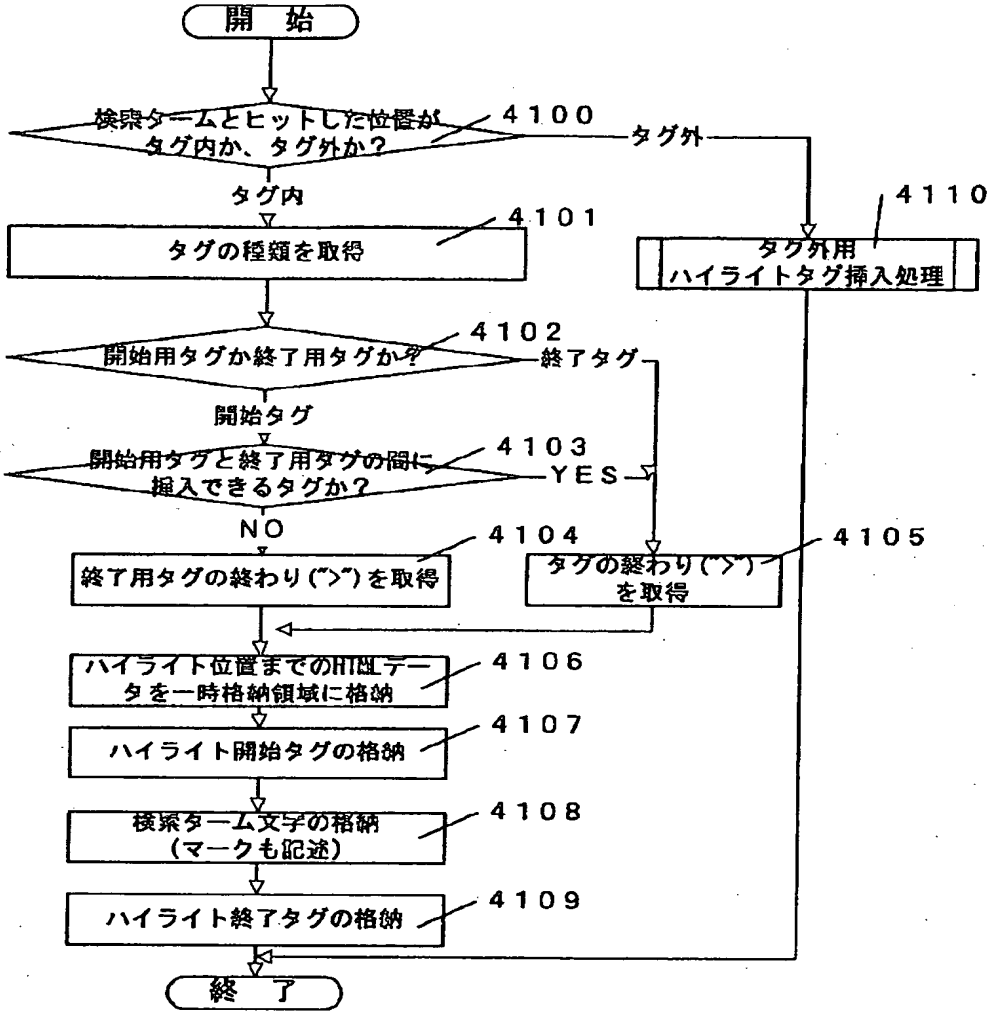
【 図４０ 】

【図４０】


開　始
ハイライト位置情報の読み出し
4000
HTML文書の一時格納領域を確保
4001
i_cnt = 0、ハイライト処理数 = 0
4002
ハイライト処理数
＜ ハイライト数
4003
NO
YES
i_cntから開始位置までのHTML文書を
HTML文書一時格納領域に格納
4004
ハイライトタグ挿入処理
4005
i_cnt = ハイライト終了タグ挿入位置
4006
i_cntからHTML文書の最後までのデータを
HTML文書一時格納領域に格納
4007
終　了

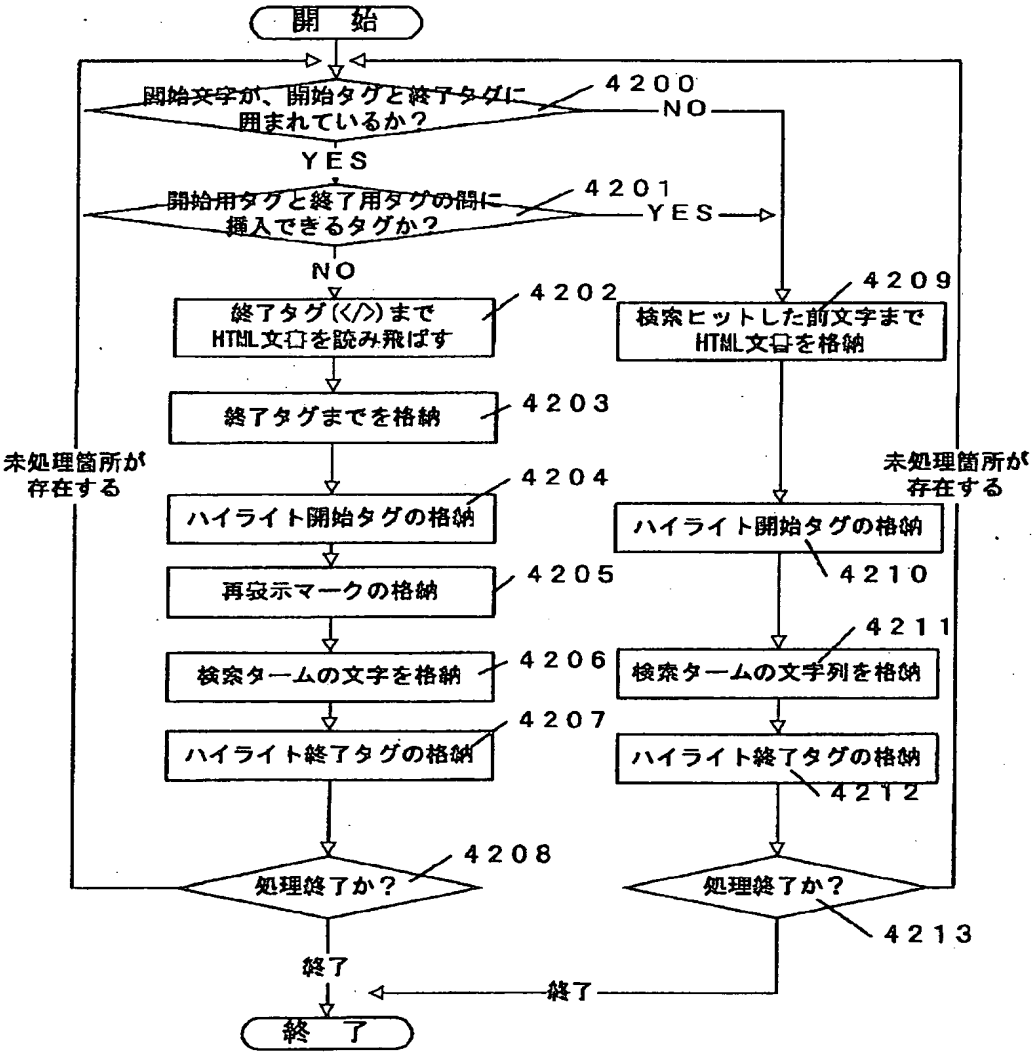
【 図4１ 】

【図４１】


開　始
検索タームとヒットした位置がタグ内か、タグ外か？
4100
タグ外
タグ内
4101
タグの種類を取得
4110
タグ外用ハイライトタグ挿入処理
4102
開始用タグか終了用タグか？
終了タグ
開始タグ
4103
開始用タグと終了用タグの間に挿入できるタグか？
YES
NO
4104
終了用タグの終わり(">")を取得
4105
タグの終わり(">")を取得
4106
ハイライト位置までのHTMLデータを一時格納領域に格納
4107
ハイライト開始タグの格納
4108
検索ターム文字の格納(マークも記述)
4109
ハイライト終了タグの格納
終　了

【図４２】

【図４２】



フロント ページの続き

(51) Int.Cl.[6]	識別記号	FI	
		G06F 15/403	380Z

(72)発明者　山崎　紀之
神奈川県戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア開発本部内

(72)発明者　青山　ゆき
神奈川県横浜市都筑区加賀原二丁目2番　株式会社日立製作所システム開発本部内